EDITED BY

JUSTIN B. BULLOCK

YU-CHE CHEN

JOHANNES HIMMELREICH

VALERIE M. HUDSON

ANTON KORINEK

MATTHEW M. YOUNG

BAOBAO ZHANG

The Oxford Handbook *of* AI GOVERNANCE

THE OXFORD HANDBOOK OF

AI GOVERNANCE

THE OXFORD HANDBOOK OF

AI GOVERNANCE

Edited by

JUSTIN B. BULLOCK,
YU-CHE CHEN,
JOHANNES HIMMELREICH,
VALERIE M. HUDSON,
ANTON KORINEK,
MATTHEW M. YOUNG,
and
BAOBAO ZHANG

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

© Oxford University Press 2024

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by license, or under terms agreed with the appropriate reproduction rights organization. Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form
and you must impose this same condition on any acquirer.

Library of Congress Cataloging-in-Publication Data
Names: Bullock, Justin B., editor.
Title: The Oxford handbook of AI governance / [edited by] Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, Baobao Zhang.
Other titles: Oxford handbook of Artificial intelligence governance
Description: New York : Oxford University Press, 2024. | Includes index. |
Identifiers: LCCN 2023035499 (print) | LCCN 2023035500 (ebook) |
ISBN 9780197579329 (hardback) | ISBN 9780197579343 (epub) |
ISBN 9780197579350
Subjects: LCSH: Artificial intelligence—Law and legislation. |
Artificial intelligence—Moral and ethical aspects. |
Artificial intelligence—Political aspects.
Classification: LCC K564.C6 O94 2024 (print) |
LCC K564.C6 (ebook) | DDC 343.09/99—dc23/eng/20230929
LC record available at https://lccn.loc.gov/2023035499
LC ebook record available at https://lccn.loc.gov/2023035500

DOI: 10.1093/oxfordhb/9780197579329.001.0001

Printed by Sheridan Books, Inc., United States of America

Contents

List of Contributors xi

Introduction 1
Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang

SECTION I: INTRODUCTION AND OVERVIEW

Justin B. Bullock

1. AI Governance: Overview and Theoretical Lenses 21
Allan Dafoe

2. AI Challenges for Society and Ethics 45
Jess Whittlestone and Samuel Clarke

3. Aligned with Whom? Direct and Social Goals for AI Systems 65
Anton Korinek and Avital Balwit

4. The Impact of Artificial Intelligence: A Historical Perspective 86
Ben Garfinkel

5. AI Governance Multi-Stakeholder Convening 109
K. Gretchen Greene

SECTION II: VALUE FOUNDATIONS OF AI GOVERNANCE

Johannes Himmelreich

6. Fairness 129
Kate Vredenburgh

7. Governing Privacy 149
Carissa Veliz

8. The Concept of Accountability in AI Ethics and Governance 164
Theodore M. Lechterman

9. Governance via Explainability 183
David Danks

10. Power and AI: Nature and Justification 198
Seth Lazar

11. AI and Structural Injustice: Foundations for Equity, Values, and Responsibility 210
Johannes Himmelreich and Désirée Lim

12. Beyond Justice: Artificial Intelligence and the Value of Community 232
Juri Viehoff

SECTION III: DEVELOPING AN AI GOVERNANCE REGULATORY ECOSYSTEM

Valerie M. Hudson

13. Transnational Digital Governance and Its Impact on Artificial Intelligence 253
Mark Dempsey, Keegan McBride, Meeri Haataja, and Joanna J. Bryson

14. Standing Up a Regulatory Ecosystem for Governing AI Decision-Making: Principles and Components 276
Valerie M. Hudson

15. Legal Elements of an AI Regulatory Permit Program 299
Brian Wm. Higgins

16. AI Loyalty by Design: A Framework for Governance of AI 320
Anthony Aguirre, Peter B. Reiner, Harry Surden, and Gaia Dempsey

17. Information Markets and AI Development 345
Jack Clark

18. Aligning AI Regulation to Sociotechnical Change 358
Matthijs M. Maas

SECTION IV: FRAMEWORKS AND APPROACHES FOR AI GOVERNANCE

YU-CHE CHEN AND MATTHEW M. YOUNG

19. The Challenge of AI Governance for Public Organizations 383
JUSTIN B. BULLOCK, HSINI HUANG, KYOUNG-CHEOL KIM, AND MATTHEW M. YOUNG

20. An Ecosystem Framework of AI Governance 398
BERND W. WIRTZ, PAUL F. LANGER, AND JAN C. WEYERER

21. Governing AI Systems for Public Values: Design Principles and a Process Framework 421
YU-CHE CHEN AND MICHAEL AHN

22. System Safety and Artificial Intelligence 441
ROEL I. J. DOBBE

SECTION V: ASSESSMENT AND IMPLEMENTATION OF AI GOVERNANCE

MATTHEW M. YOUNG AND YU-CHE CHEN

23. Assessing Automated Administration 461
CARY COGLIANESE AND ALICIA LAI

24. Transparency's Role in AI Governance 479
ALEX INGRAMS AND BRAM KLIEVINK

25. The Anatomy of AI Audits: Form, Process, and Consequences 495
INIOLUWA DEBORAH RAJI

26. Mitigating Algorithmic Biases through Incentive-Based Rating Systems 517
NICOL TURNER LEE

27. Role and Governance of Artificial Intelligence in the Public Policy Cycle 534
DAVID VALLE-CRUZ AND RODRIGO SANDOVAL-ALMAZÁN

SECTION VI: AI GOVERNANCE FROM THE GROUND UP (VIEWS FROM THE PUBLIC, IMPACTED COMMUNITIES, AND ACTIVISTS WITHIN THE TECH COMMUNITY)

BAOBAO ZHANG

28. Public Opinion Toward Artificial Intelligence 553
BAOBAO ZHANG

29. Adding Complexity to Advance AI Organizational Governance Models 572
JASMINE MCNEALY

30. The Role of Workers in AI Ethics and Governance 584
NATALIYA NEDZHVETSKAYA AND J. S. TAN

31. Structured Access: An Emerging Paradigm for Safe AI Deployment 604
TOBY SHEVLANE

32. AI, Complexity, and Regulation 619
LAURIN B. WEISSINGER

SECTION VII: ECONOMIC DIMENSIONS OF AI GOVERNANCE

ANTON KORINEK

33. Technological Unemployment 641
DANIEL SUSSKIND

34. Harms of AI 660
DARON ACEMOGLU

35. AI and the Economic and Informational Foundations of Democracy 707
CARLES BOIX

36. Governing AI to Advance Shared Prosperity 726
KATYA KLINOVA

37. Preparing for the (Non-Existent?) Future of Work 746
ANTON KORINEK AND MEGAN JUELFS

SECTION VIII: DOMESTIC POLICY APPLICATIONS OF AI

JOHANNES HIMMELREICH

38. Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance 779
KURT GLAZE, DANIEL E. HO, GERALD K. RAY, AND CHRISTINE TSANG

39. Watching the Watchtower: A Surveillance AI Analysis and Framework 797
STEPHEN CAINES

40. Smart City Technologies: A Political Economy Introduction to Their Governance Challenges 820
BEATRIZ BOTERO ARCILA

41. Artificial Intelligence in Healthcare 838
NAKUL AGGARWAL, MICHAEL E. MATHENY, CARMEL SHACHAR, SAMANTHA WANG, AND SONOO THADANEY-ISRANI

42. AI, Fintech, and the Evolving Regulation of Consumer Financial Privacy 860
NIKITA AGGARWAL

SECTION IX: INTERNATIONAL POLITICS AND AI GOVERNANCE

JUSTIN B. BULLOCK

43. Dueling Perspectives in AI and U.S.–China Relations: Technonationalism vs. Technoglobalism 881
JEFFREY DING

44. Mapping State Participation in Military AI Governance Discussions 895
JUSTIN KEY CANFIL AND ELSA B. KANIA

45. AI, the International Balance of Power, and National Security Strategy 914
MICHAEL C. HOROWITZ, SHIRA PINDYCK, AND CASEY MAHONEY

46. The Ghost of AI Governance Past, Present, and Future: AI Governance in the European Union 937
Charlotte Stix

47. AI and International Politics 959
Amelia C. Arsenault and Sarah E. Kreps

48. The Critical Roles of Global South Stakeholders in AI Governance 981
Marie-Therese Png

49. NATO's Role in Responsible AI Governance in Military Affairs 1015
Zoe Stanley-Lockman and Lena Trabucco

Index 1043

Contributors

Daron Acemoglu, Institute Professor, Department of Economics, MIT

Nakul Aggarwal, MD-PhD Candidate, Medical Scientist Training Program, University of Wisconsin-Madison

Nikita Aggarwal, Lecturer in Law, UCLA School of Law

Anthony Aguirre, Faggin Professor of the Physics of Information, University of California at Santa Cruz

Michael Ahn, Associate Professor, Department of Public Policy and Public Affairs, McCormack Graduate School, University of Massachusetts-Boston

Beatriz Botero Arcila, Assistant Professor of Law, Sciences Po

Amelia C. Arsenault, PhD student, Department of Government, Cornell University, Cornell University

Avital Balwit, Research Scholar, Future of Humanity Institute, Oxford University

Carles Boix, Robert Garrett Professor of Politics and Public Affairs, Princeton University

Joanna J. Bryson, Professor of Ethics and Technology, Centre for Digital Governance, Hertie School, Berlin, Germany

Justin B. Bullock, Associate Professor Affiliate Evans School of Public Policy and Governance, University of Washington

Stephen Caines, Deputy Chief Innovation Officer, City of San Jose

Justin Key Canfil, Assistant Professor, Institute for Politics and Strategy, Carnegie Mellon University

Yu-Che Chen, Isaacson Professor, School of Public Administration, University of Nebraska at Omaha

Jack Clark, Co-Founder of Anthropic; Co-Chair of the AI Index Steering Committee, Stanford Institute for Human-Centered Artificial Intelligence (HAI)

Samuel Clarke, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Cary Coglianese, Edward B. Shils Professor of Law and Political Science, University of Pennsylvania

Allan Dafoe, Senior Staff Research Scientist, DeepMind

David Danks, Professor of Data Science & Philosophy, University of California, San Diego

Gaia Dempsey, CEO of Mataculus, Santa Monica, CA

Mark Dempsey, Senior Policy Advisor (consultant), Hertie School of Governance (Executive MPA graduate 2021)

Jeffrey Ding, Assistant Professor of Political Science at George Washington University

Roel I. J. Dobbe, Assistant Professor of Data-driven and Algorithmic Systems, Delft University of Technology

Ben Garfinkel, Research Fellow, Faculty of Philosophy, University of Oxford

Kurt Glaze, Senior Advisor, U.S. Social Security Administration

K. Gretchen Greene, Harvard Kennedy School and The Hastings Center

Meeri Haataja, CEO & Co-Founder, Saidot, Helsinki, Finland

Brian Wm. Higgins, Partner, Blank Rome LLP

Johannes Himmelreich, Assistant Professor in Public Administration and International Affairs, Maxwell School of Citizenship & Public Affairs at Syracuse University

Daniel E. Ho, William Benjamin Scott and Luna M. Scott Professor of Law; Professor of Political Science; Senior Fellow, Stanford Institute for Economic Policy Research, Stanford University

Michael C. Horowitz, Director of Perry World House and Richard Perry Professor at the University of Pennsylvania

Hsini Huang, Assistant Professor, Leiden University

Valerie M. Hudson, University Distinguished Professor, Department of International Affairs, The Bush School of Government and Public Service, Texas A&M University

Alex Ingrams, Assistant Professor, Institute of Public Administration, Leiden University

Megan Juelfs, Senior Associate Director for Research, Office of Research Services, University of Virginia Darden School of Business

Elsa B. Kania, Adjunct Senior Fellow with the Technology and National Security Program at the Center for a New American Security

Kyoung-Cheol Kim, PhD student/Instructor, Department of Public Administration and Policy, The University of Georgia

Bram Klievink, Professor of Public Policy & Digitalisation, Public Administration Institute, Leiden University

Katya Klinova, Head of AI, Labor and the Economy, Partnership on AI

Anton Korinek, Professor, Department of Economics and Darden School of Business, University of Virginia

Sarah E. Kreps, John L. Wetherill Professor and Director of the Tech Policy Institute at Cornell University

Alicia Lai, University of Pennsylvania Carey Law School

Paul F. Langer, Postdoctoral Researcher, German University of Administrative Sciences Speyer

Seth Lazar, Professor of Philosophy, Australian National University

Theodore M. Lechterman, Assistant Professor of Philosophy at IE University

Nicol Turner Lee, Senior Fellow, Center for Technology Innovation, the Brookings Institution

Désirée Lim, Research Associate in the Rock Ethics Institute, Assistant Professor of Philosophy, Penn State

Matthijs M. Maas, Senior Research Fellow, Legal Priorities Project

Casey Mahoney, Department of Political Science, University of Pennsylvania

Michael E. Matheny, Professor of Biomedical Informatics, Vanderbilt University Medical Center

Keegan McBride, Postdoctoral Researcher, Centre for Digital Governance, Hertie School

Jasmine McNealy, Associate Professor, University of Florida

Nataliya Nedzhvetskaya, PhD Candidate, Department of Sociology, University of California, Berkeley

Shira Pindyck, Postdoctoral Fellow in Technology and International Security, UC Institute on Global Conflict and Cooperation and the Livermore and Los Alamos National Laboratories

Marie-Therese Png, DeepMind Scholar, Oxford Internet Institute

Inioluwa Deborah Raji, Fellow, Mozilla Foundation

Gerald K. Ray, US Social Security Administration, Retired

Peter B. Reiner, Professor of Neuroethics, Department of Psychiatry, University of British Columbia

Rodrigo Sandoval-Almazán, Associate Professor, Universidad Autonoma del Estado de Mexico

Carmel Shachar, Assistant Clinical Professor of Law and Faculty Director, Center for Health Law and Policy Innovation, Harvard Law School, Harvard Law School

Toby Shevlane, Research Scientist, AGI Strategy & Governance Team, Google DeepMind

Zoe Stanley-Lockman, Associate Research Fellow in the Military Transformations Programme at the Institute of Defence and Strategic Studies at the S. Rajaratnam School of International Studies

Charlotte Stix, University of Technology

Harry Surden, Professor of Law, University of Colorado

Daniel Susskind, Senior Research Associate, Institute for Ethics in AI, Oxford University

J. S. Tan, PhD student in the International Development Group at DUSP, MIT

Sonoo Thadaney-Israni, Executive Director, PRESENCE @ Stanford Medicine, Stanford University

Lena Trabucco, Visiting Research Scholar, Stockton Center for International Law, US Naval War College

Christine Tsang, Executive Director, Stanford Regulation, Evaluation, and Governance Lab, Stanford University

David Valle-Cruz, Professor of Applied Artificial Intelligence and Data Science, Universidad Autónoma del Estado de México

Carissa Veliz, Associate Professor at the Institute for Ethics in AI, University of Oxford

Juri Viehoff, Lecturer in Political Theory, University of Manchester

Kate Vredenburgh, Assistant Professor, Department of Philosophy, Logic, and Scientific Method, The London School of Economics

Samantha Wang, Clinical Assistant Professor, Department of Medicine, Stanford University

Laurin B. Weissinger, Cybersecurity Fellow, Yale Law School, Yale University

Jan C. Weyerer, Research Associate, Chair for Information and Communication Management, German University of Administrative Sciences Speyer

Jess Whittlestone Head of AI Policy at the Centre for Long-Term Resilience

Bernd W. Wirtz holds the Chair of Information and Communication Management at the German University of Administrative Sciences Speyer, Germany.

Matthew M. Young, Assistant Professor of Public Administration, Faculty of Governance and Global Affairs, Leiden University

Baobao Zhang, Assistant Professor, Political Science, Syracuse University

INTRODUCTION

JUSTIN B. BULLOCK, YU-CHE CHEN, JOHANNES HIMMELREICH, VALERIE M. HUDSON, ANTON KORINEK, MATTHEW M. YOUNG, AND BAOBAO ZHANG

Artificial Intelligence (AI), especially in the form of modern machine learning techniques, has exploded in both its capabilities and its integration into society in recent years. It has begun to significantly influence humans and the operation of corporations, markets, governments, and society more generally, making it necessary to examine the myriad impacts of AI, both positive and negative. As with other powerful technologies, it is important that AI is carefully and deliberately governed by society. This is the domain of AI governance.

Governance can be viewed as the process by which societal activity is organized, coordinated, steered, and managed. Furthermore, governance also refers to the set of norms, rules, regulations, institutions, and processes that coordinate the actions and interactions of otherwise uncoordinated actors to achieve outcomes that are jointly desirable for them; that is, outcomes that internalize externalities and recognize the role of public goods.

Long ago, governance simply referred to the process of coordinating groups of humans, for example, collectives such as tribes that consisted of individual humans, illustrated in panel (a) of Figure 1.1. Over time, aided by new forms of technology that allowed for the processing of more and more information and improved means of communication across humans, societies became more complex, and thus more complex governance entities and mechanisms were needed. These entities evolved into governments, organizations, and markets and they serve core functions in modern systems of governance—we will henceforth refer to them as artificial entities to distinguish them from humans. These artificial entities deal with the complexity of their environment in ways that allow humans to expand their group capacities. Governments developed new tools to both coordinate and control humans at a scale that was impossible before their creation. Private organizations, such as corporations, made it possible to provide services, goods, and knowledge at scale. Both governments and private organizations increasingly relied on bureaucracies, allowing them to coordinate large groups of humans towards a common goal in a standardized and hierarchical manner. Moreover, markets increased in their scope and reach, efficiently compiling information on the scarcity of goods and services into prices that guided the allocation of resources and greatly improved human standards of living. In short, the new artificial

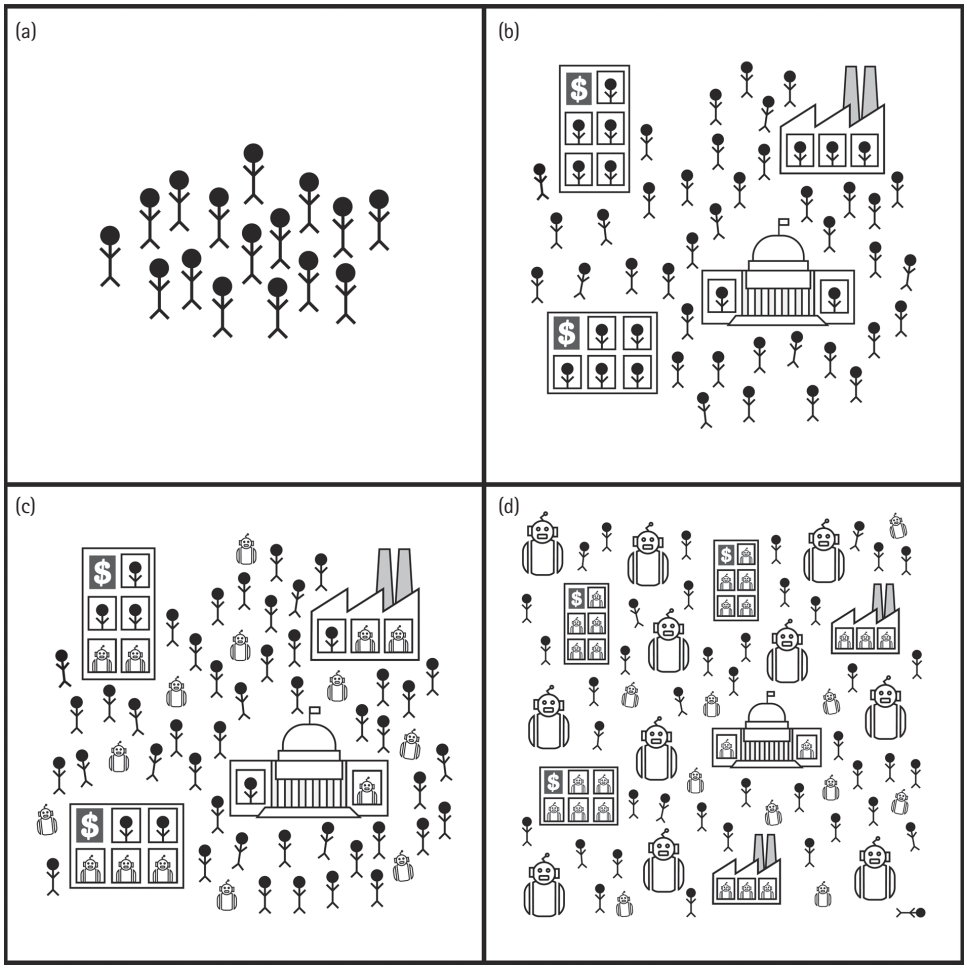

FIGURE 1.1 Evolution of the governance ecosystem

Copyright (c) 2023 by Justin Bullock, Anton Korinek, and Davis Taliaferro under a CC BY 4.0 License.

entities greatly advanced the human condition. The resulting governance ecosystem is illustrated in panel (b) of Figure 1.1.

It soon became clear that there were times when the new artificial entities acted as if they had interests of their own that conflicted with their human creators and held human society back from reaching the full potential of human flourishing. It seemed as if artificial entities increasingly acquired agency of their own, independently pursuing goals, influencing the allocation of resources, and wielding real power. For example, the bureaucracies within governments and corporations did not always serve the interest of their stakeholders, suffering from drawbacks such as rigidity, dehumanization, and organizational politicking. Likewise, markets also came with externalities that imposed significant inefficiencies and costs on individual humans, groups, and societies.

The growing influence, prominence, and power of these artificial entities thus posed increasingly complex governance challenges even before the advent of AI. These challenges

included (1) how to develop and implement governance mechanisms to allow humans to steer, direct, and control the artificial entities, and (2) how to integrate the artificial entities into the governance ecosystem more broadly, i.e., how to shape the interactions among entities, their access to resources, their power relationships, and the ways in which they process information. In short, we needed to understand both how to govern the artificial entities themselves and how to integrate them into the governance ecosystem.

AI systems are the latest set of entities to join the governance ecosystem and give rise to new challenges, as illustrated in panel (c) of Figure 1.1. The goal of AI Governance is to integrate the new and fast evolving AI systems into our framework of governance. AI systems present similar governance challenges to other artificial entities: (1) how to ensure that AI is steered, managed, and controlled to the benefit of human society as opposed to its detriment; and (2) how AI systems reshape the interactions of existing actors in the governance ecosystem to further human interests. While there are some similarities between how AI and traditional artificial entities affect our system of governance, the ways in which the challenges of AI governance manifest themselves are unique to AI. The goal of this Handbook is to provide a systematic examination of these challenges.

AI systems differ from existing artificial entities in two significant ways that elevate the importance of AI governance. The first is the fast pace of progress in AI, and the second is the capacity of AI to increasingly make decisions without humans in the loop. Starting with the first difference, the pace of progress in AI, or one might say the pace of evolution of AI entities, vastly outstrips the pace of evolution of humans and other artificial entities in the governance ecosystem. Driven in part by Moore's Law and by an ever-growing amount of resources pouring into the field, the capabilities of AI systems are progressing relentlessly. This creates difficulties for understanding AI systems well enough to govern and control them and for managing their interactions throughout society, both when they interact with individual humans and when they are deployed in the service of other artificial entities.

The second difference arises because the core constitutive parts of AI systems, once they are in operation, are not humans but computing machines. Traditional artificial entities, such as governments or corporations, still have humans in the loop for making most important decisions, making it easier to decipher their decisions in a transparent manner, to correct mistakes, and to control them. In a sense, the humans that act on behalf of artificial entities provide a robustness check that makes it easier to avoid failure modes and grossly unethical behaviors. By contrast, AI systems make decisions independently once they are in operation, making it even more important that their design is in alignment with human values. Moreover, AI systems play an increasingly important role in the operation of all the other artificial entities; for example, in financial markets, only a small sliver of all transactions is executed by humans. As a result, they leave a footprint on all aspects of the governance ecosystem.

These two differences taken together give rise to what may be the ultimate challenge for AI Governance: to prepare for the possibility that ever-more advanced AI systems that operate independently may achieve levels of intelligence and by extension capabilities that are beyond human comprehension and direct human control, giving them an ever-greater role in the governance ecosystem. This is depicted in panel (d) of Figure 1.1. Navigating this transition and ensuring that the resulting new governance ecosystem still serves the interests of humans may well be one of the greatest challenges for humanity in the twenty-first century.

Overview of *The Oxford Handbook of AI Governance*

This Handbook is a concerted effort to bring together the leading experts on AI Governance and compile the emerging new body of knowledge on the topic. We identified the leading voices from an interdisciplinary and diverse set of backgrounds and encouraged them to write about the topics they viewed as the most pressing to improve the state of dialogue and practice around AI Governance. Because modern AI systems outstretch the boundaries of nation states, we enlisted a set of scholars from all around the globe. Our ultimate goal is to rise to the challenge of governing AI and integrating AI into the broader ecosystem of governance. This is reflected throughout the 49 chapters of this handbook, which are split into the following nine sections.

Section I is an "Introduction and Overview" section, which lays out the myriad issues and challenges concerning the field of AI Governance from several complementary perspectives. Section II, "Value Foundations of AI Governance," turns to the needed ethical and value foundations for understanding the directions in which we should want to steer the development, use, and application of advanced AI systems. With this direction in mind, Section III, "Developing an AI Governance Regulatory Ecosystem," turns to an examination of the regulatory tools, processes, and structures that can assist in steering AI systems effectively and democratically. This leads naturally to Section IV, "Frameworks and Approaches for AI Governance," which examines the various frameworks and approaches for directing and controlling AI and how it interacts with other entities throughout the governance ecosystem, building on the fields of governance, public administration, and public management. Continuing with the use of these disciplines, Section V, "Assessment and Implementation of AI Governance," looks more carefully at how to assess and implement the uses of AI and its relationship to governance. The remainder of the Handbook takes specific areas of concern and opportunity for AI Governance and closely examines them. Section VI, "AI Governance from the Ground Up," takes a bottom up perspective to explore voices that are often neglected in structuring both the development and use of AI systems and governance systems. Section VII, "Economic Dimensions of AI Governance," analyzes the challenges that transformative advances in AI may bring in economic areas such as income distribution and unemployment. Section VIII, "Domestic Policy Applications of AI," describes the range of contexts in which AI systems are used to enhance governance in a domestic setting. Finally, Section IX, "International Politics and AI Governance," explores AI Governance through the lens of international politics.

The collected chapters within these nine sections provide carefully reasoned arguments, language, systematization, formal modeling, and empirical observation to guide the reader through the complex governance opportunities and challenges presented by the development and deployment of AI systems. We are left with few easy answers to these governance challenges. It is our hope that the dialogue that is captured by this Handbook will help to guide and provide pathways and knowledge that help humanity to collectively govern AI and control the AI governance systems that we are currently creating and deploying.

In the following, we provide brief summaries of each of the 49 chapters, organized by section. This serves to provide readers with an overview of the Handbook and with directions on which chapters are most relevant to their specific areas of interest.

Section I: Introduction and Overview

Section I highlights many of the core challenges AI Governance faces. The chapters include a deliberate overview of the field, an examination of the challenges for society and ethics, a formal treatment of the alignment problem, a historical perspective, and a practical and pragmatic guide to convening various stakeholders together to improve AI Governance. Taken together, this section gives the reader a sense of scope of the challenges and several approaches for making sense of this vast and challenging landscape.

Chapter 1, from Allan Dafoe, is titled "AI Governance: Overview and Theoretical Lenses." This chapter gives us an early working definition of AI Governance and then discusses many of the challenges to effective AI Governance that is beneficial for humanity and avoids extreme risks. Dafoe highlights the particular challenges of AI governance in the domain of great powers competition, a topic revisited later in multiple chapters. Dafoe argues that much more work on AI Governance is needed, particularly "the problem of devising global norms, policies, and institutions to promote the beneficial development and use of advanced AI."

Chapter 2, from Jess Whittlestone and Samuel Clarke, is titled "AI Challenges for Society and Ethics." This chapter highlights that the current and growing use of AI in sectors such as healthcare, finance, and policing, while providing tools that may be highly beneficial to society, also presents complex challenges and risks to society and ethics. Whittlestone and Clarke note three important categories of benefits and opportunities presented by AI including: (1) improving the quality and length of people's lives, (2) improving our ability to tackle problems as a society, and (3) enabling moral progress and cooperation. However, they also note five considerable harms and risks that AI presents including: (1) increasing the likelihood or severity of conflict, (2) making society more vulnerable to attack or accident, (3) increasing power concentration, (4) undermining society's ability to solve problems, and (5) losing control of the future to AI systems. In closing they argue that AI Governance "should be focused on identifying and implementing mechanisms which enable benefits and mitigate harms of AI," and that AI Governance needs to develop methods that "improve our ability to assess and anticipate the impacts of AI; and to make decisions even in the face of normative uncertainty and disagreement."

Chapter 3, from Anton Korinek and Avital Balwit, is titled "Aligned with Whom? Direct and Social Goals for AI Systems." This chapter explores the AI alignment problem, one of the core challenges of AI Governance. Korinek and Balwit argue that when considering AI alignment, we should be concerned about both direct alignment "whether an AI system accomplishes the goals of the entity operating it" and social alignment "the effects of an AI system on larger groups or on society more broadly." Using this language and the framework of externalities, Korinek and Balwit distinguish different characteristics of direct and social alignment and highlight that social alignment, in particular, requires both enforcing

existing norms on AI developers and operators and designing new norms that apply directly to AI systems.

Chapter 4, from Ben Garfinkel, is titled "The Impact of Artificial Intelligence: A Historical Perspective." This chapter examines AI as akin to a general purpose technology that may become a revolutionary technology. Garfinkel argues that historically general purpose technologies have, gradually, led to transformations in economies, militaries, and politics. However, if AI does generally supplant human labor across many domains and sectors, it may then be regarded as a revolutionary technology to be considered in impact akin to the Neolithic Revolution and the Industrial Revolution.

The final chapter of this section is Chapter 5. This chapter is from Gretchen Greene, and it is titled "AI Governance Multi-Stakeholder Convening." In this chapter, Greene provides a personal reflection and general advice on AI ethics and governance from their work bringing together AI Governance stakeholders for a multi-disciplinary, multi-stakeholder collaboration. Greene offers a blueprint for success in bringing together foundational knowledge so that technical and non-technical experts alike can bring their expertise and concern to the governance process. It also includes a curated list of 42 questions that can serve as guides for convening diverse stakeholders around questions and challenges of AI Governance.

Section II: Value Foundations of AI Governance

Section II stands in the conviction that all governance problems raise value questions. Even if a governance problem itself does not consist in a value question, theories of values—theories concerning what *should* be done and why—can deepen our understanding of the governance problem and potential disagreement over it. Value foundations of AI Governance are thus both substantive—they are foundations on which governance structures and policies can be built—and analytical, that is, foundations that help to understand and capture a range of viewpoints, concerns, and experiences. The section is roughly organized by going from the more specific to more general topics. In each chapter, the authors ask what a certain value or concept means, they make distinctions, relate the concept to AI governance, and discuss upshots of their discussion.

Chapter 6 is on fairness. Fairness is perhaps the value that is invoked most prominently in discussions of AI Governance. However, what is often overlooked is that fairness as such is narrow and specific. Fairness is only one of many ingredients of justice: "In situations of serious and pervasive injustice," Kate Vredenburgh argues, "fairness has no value." Vredenburgh surveys the existing discussions of fairness, contrasts fairness and justice, and suggests five policy avenues to make AI Governance fair. This chapter stands to reorient the debate on fairness in AI. When "fairness" is invoked, often justice is meant.

Chapter 7 is on privacy. Carissa Veliz writes from the perspective of moral philosophy. She first explains what this moral right to privacy is—a distinction between access and control theories of privacy is important here—and then makes explicit why privacy is valuable. Privacy is not only about preventing individual harms but also about securing collective goods, such as democracy. In a nutshell, we could say: Privacy is about power.

These considerations lead to an important upshot: Consent, in contrast to the prominent place it often enjoys in practice, plays only a very limited role in securing privacy.

Chapter 8 turns to accountability. Accountability is not only taken to be a central value of governance and administration, but some suggest that we should aim for "accountable AI." Ted Lechterman analyzes what "accountability" means. He argues that accountability is, in a sense, only a derivative, a procedural, and not itself a primary or substantive value. "Accountability's primary job description is to verify compliance with substantive normative principles—once those principles are settled." So, the question is not so much *whether* there should be accountability but accountability in light of *what*? Demanding accountability hence demands far less than is often assumed.

Chapter 9 is about explainability. David Danks argues that explainable AI (XAI) is an important tool of AI Governance. However, the question of what it means for an AI to be "explainable" poses several challenges. Danks distinguishes different theories of what an explanation is—there is, for example, a trade-off between an explanation being true and it being intuitive—and to whom such explanations are given. This conceptual background is important to specify the goals of AI Governance accurately. Otherwise, we risk "that explanations provided by XAI are not the kind required for governance."

Chapter 10 is about power. Seth Lazar gives a definition of "having power over someone." (Just so much: The simple analysis fails that *A* has power over *B* when *A* can get *B* to do something that they would otherwise not do.) Lazar then argues that power is perhaps the central problem in AI Governance. Instead of concentrating on the aims or goals for which AI is used, the question of power asks not for which end power is exercised but by whom. The bigger picture here is that this chapter highlights how political philosophy can contribute to AI Governance. Power raises the issues of justification, legitimacy, and authority—three concepts that need to be clearly distinguished, each of which is a central research topic in political philosophy.

Chapter 11 is about structural injustice. This chapter continues lines of inquiry that earlier chapters opened but didn't pursue. From the chapter on power, which concentrates on power between agents, this chapter picks up the idea that social structures exert power. From the chapter on fairness, this chapter picks up that justice should be a fundamental concern of AI Governance. Désirée Lim and Johannes Himmelreich argue that the conceptual tool of structural injustice is indispensable for good AI Governance. They distinguish different ways in which AI can be involved in structural injustice and outline a theory of structural injustice, that is, a theory of justice that places heavy emphasis on social structural explanations. This approach to normative AI Governance, or so the authors argue, is preferable over alternative approaches that focus instead on harms and benefits, or mere statements of values.

Chapter 12 is about the values of AI Governance beyond justice. This chapter, in turn, picks up where the chapter on structural injustice has left off. Juri Viehoff explains what the value of community is and argues that it is a value that is of high importance in AI Governance. Community fills a conceptual and analytical gap. It highlights concerns that are not adequately captured by the ideas of fairness, explainability, privacy, or justice. The problems of a lack of community can already be felt today. AI is bound to make matters worse. Viehoff systematically identifies pathologies of decommunitarization, driven by automation, datafication, and the disappearing public sphere. The idea of community is hence

an important analytical lens to understand shortcomings of our status quo as well as risk of AI in the future.

On a personal note, the authors of this section wish to commemorate the life and work of Waheed Hussain. Hussain had planned to contribute a chapter to this section. His work has been an inspiration for many of us. His passing leaves voids large and small.

Section III: Developing an AI Governance Regulatory Ecosystem

Section III is titled "Developing an AI Governance Regulatory Ecosystem." This section has six chapters that examine various facets of regulating AI systems. The section examines transnational approaches, ecosystem components, permitting, loyalty, information markets, and sociotechnical changes. The section provides several examinations on how to set up the regulatory ecosystem by which we govern the use of AI.

Chapter 13, from Mark Dempsey, Keegan McBride, Meeri Haataja, and Joanna J. Bryson, is titled "Transnational Digital Governance and Its Impact on Artificial Intelligence (AI)." In this chapter the authors examine and discuss the European Union's approach to transnational AI regulation. They argue that the EU commission's approach, given the "Brussels effect," has a large global impact on the AI regulatory space. The authors explore key EU documents, laws, approaches to regulating AI, and discuss what the rest of the world might learn from this approach.

Chapter 14, from Valerie Hudson, is titled "Standing Up a Regulatory Ecosystem for Governing AI Decision Making: Principles and Components." In this chapter Hudson identifies the core components necessary for a governance regime of AI decision-making. Hudson explores how basic human rights to know, to appeal, and to litigate the systems they interact with leads to a governance regime that would include standard setting, training, insurance, procurement, identification, archiving, and testing, among other functions. Hudson explores the regulatory ecosystem and examines governance ecosystem questions such as how "to render such governance both robust and sustainable over time? How could checks and balances be woven into the very structure of that ecosystem, so that it remains functional, sustainable, and adaptive?"

Chapter 15, from Brian Wm Higgins, is titled "Legal Elements of an AI Regulatory Permit Program." Higgins argues for a specific risk mitigation strategy for AI deployment. Higgins argues for a permitting strategy "whereby government administrators act as market gatekeepers and allow AI systems to operate in commerce only if they meet specific technical and other standards, and their owners demonstrate they can operate their systems within acceptable levels of risk while also maintaining compliance with individualized permit terms and conditions." Higgins compares and contrasts this approach with the approaches taken for "AI-based medical devices and the European Commission's proposed regulatory framework for AI." Higgins' approach here provides another pathway by which AI companies can be regulated in such a way as to protect human rights and the public interest.

Chapter 16, from Anthony Aguirre, Peter B. Reiner, Harry Surden, and Gaia Dempsey, is titled "AI Loyalty by Design: A Framework for Governance of AI." These authors argue that

there is often an inherent tension between the creators of AI systems and end users of those systems. As AI is deployed throughout society it is being used in domains where loyalty to the end user is important to the end user. This includes domains such as healthcare and the legal system where conflicts of interest arise between the AI developers and the AI users. At its core, the authors consider AI loyalty as a "principle that AI systems should be designed, from the outset, to primarily and transparently benefit their end users, or at minimum clearly communicate conflict-of-interest tradeoffs, if they cannot be eliminated."

Chapter 17 is from Jack Clark. It is titled "Information Markets and AI Development." For this chapter, Clark argues that to provide more capacity for governments to effectively regulate and govern AI an intervention is needed. This intervention focuses on using the metrics and measures of the AI research community as informal devices of self-regulation that also point to measures that can be used by governments and their policymaking apparatus to steer the development and deployment of AI by both the private sector and academia.

Chapter 18 is the final chapter for Section III. The chapter is titled "Aligning AI Regulation to Sociotechnical Change," and it is authored by Matthijs M. Maas. In this chapter Maas argues that the AI regulatory ecosystem should move past the silos of law-centric and technology-centric regulatory approaches and instead emphasize "how, when, and why AI applications enable patterns of 'sociotechnical change.'" Maas argues that the focus on sociotechnical change "can help analyze when and why AI applications actually do create a meaningful rationale for new regulation—and how they are consequently best approached as targets for regulatory intervention." Maas explores six problem logics for AI regulation including (1) ethical challenges, (2) security risks, (3) safety risks, (4) structural shifts, (5) public goods, and (6) governance disruption. Maas concludes by using the six problem logics to "improve the regulatory triage, tailoring, timing & responsiveness, and regulatory design of AI policy."

Section IV: Frameworks and Approaches for AI Governance

The existing literature on AI Governance calls for more theoretical, integrated, and broad-based frameworks and approaches to understanding AI Governance. Collectively, the four chapters in the section on Frameworks and Approaches for AI Governance answer the call with the focus on advancing public values. This section begins with chapter 19 by Justin B. Bullock, Hsini Huang, Kyoung-Cheol Kim, and Matthew M. Young to ground our inquiry in theoretical insights into the challenges of AI governance in public sector organizations. The next chapter, chapter 20, by Bernd W. Wirtz, Paul F. Langer, and Jan C. Weyerer provides an integrated multi-level broad-based framework to address these fundamental challenges of AI governance. Chapter 21 by Yu-Che Chen and Michael Ahn advances AI governance by articulating phases of AI governance and their integration. Roel I. J. Dobbe's chapter, Chapter 22 offers a broad-based interdisciplinary approach to address AI safety governance challenges. Systems thinking is either explicit or embedded in all these chapters in their development of AI governance frameworks and approaches.

Justin B. Bullock, Hsini Huang, Kyoung-Cheol Kim, and Matthew M. Young, in "The Challenge of AI Governance for Public Organizations," provide an informative theoretical

development for identifying the challenge of AI governance for public organizations. They build on the seminal works of Max Weber on bureaucracy and Herbert Simon on administrative behavior. The integration of macro (Weber) and micro (Simon) perspectives provides integrated theoretical insights into AI governance in public organizations. The discussion about Simon's illumination of the respective roles of humans and machines in public organizations provides a theoretical basis for AI governance. Moreover, this chapter offers key factors for consideration in developing AI governance solutions. These factors include, as the authors state, "bounded rationality, hierarchy, facts, specialization, and human judgment," as well as the dynamics of their inter-relations.

"An Ecosystem Framework of AI Governance," by Bernd W. Wirtz, Paul F. Langer, and Jan C. Weyerer, offers an integrated and broad-based framework of AI governance. Such a framework identifies AI governance challenges and opportunities. It also offers solutions in the form of guidelines, activities, and policies. The underlying approach is an ecosystem, one that recognizes and models multiple layers and dynamics within and between layers. These layers include AI systems, AI governance challenges, AI multi-stakeholder governance process, AI governance mechanisms, and AI governance policy. The multi-stakeholder governance process is broad-based and inclusive in engaging government, industry, academia, and civil society in the dynamics of framing, assessment, evaluation, and management.

"Governing AI Systems for Public Values," by Yu-Che Chen and Michael Ahn, further enriches the development of AI governance frameworks and approaches by providing a principle-based process-oriented design. The principles are human-centered, stakeholder-focused, and lifecycle-scoped. The process approach includes phases of (1) goal setting; (2) iterative development decisions on data, models, and results; (3) decisions on public service; and (4) the assessment of the impacts. Additionally, this approach includes the application of the three proposed governance principles in specific phases. This process-oriented framework integrates actions at various phases into an integral whole and underscores the importance of transparency. The focus on individual AI systems makes the recommendations actionable.

Roel I. J. Dobbe's chapter, "System Safety and Artificial Intelligence," uses a systems perspective to understand the potential harms of AI systems and ways to mitigate them. This chapter provides an in-depth treatment of systems concepts applied to AI systems. More specifically, Dobbe illustrates the applicability of the seven lessons drawn from seminal work on safety systems to AI systems. One key lesson is to consider the broader socio-technical context of AI system design, including the use context, stakeholders, and institutional environment. Another one speaks to the importance of building an organizational culture to achieve safety goals. Moreover, this chapter also calls for a transdisciplinary approach to governing AI systems for safety. Such an approach underscores the need for an integrated broad-based framework for effective AI governance.

Section V: Assessment and Implementation of AI Governance

Section V is titled "Assessment and Implementation of AI Governance." It contains five contributions that examine the ways in which AI may be used by governments and the

ways in which the AI systems themselves need to be better governed if they are to be used by governments.

Chapter 23, from Cary Coglianese and Alicia Lai, is titled "Assessing Automated Administration." This chapter examines whether and when it is appropriate for public administrators to use AI to automate government tasks. The chapter highlights that human administrators have strengths and weaknesses of their own, and that these characteristics make these administrators very well suited for some tasks. For other tasks, AI systems may improve upon the status quo of government decision making. However, before the deployment of AI systems is appropriate, the authors argue that four preconditions should be met: (1) adequate resources, (2) goal clarity and precision, (3) data availability, and (4) external validity. In addition to these preconditions, the value of the system should also be clear. That is, it should be clear that the AI system will be an improvement over the status quo. To examine the value of the AI system the authors encourage administrators to examine: (1) task performance, (2) user or beneficiary impacts, and (3) societal impacts. Successful AI Governance will only occur if administrators are diligent in how AI systems are deployed by governments.

Chapter 24 is from Alex Ingrams and Bram Klievink. Their chapter is titled "Transparency's Role in AI Governance." This contribution to the Handbook does a deep dive into the issue of transparency. Transparency is often considered an important element to good governance. The authors' review the current debates around governance and transparency and also across AI and transparency. From this review they identify four main approaches for delivering transparency. These approaches are (1) constructivist, (2) democratization, (3) legal, and (4) capacity building. The authors argue that instead of thinking of these approaches as in conflict, they can be integrated across whether they emphasize individuals or institutions and whether they seek transparency that is accessible to the public or to experts. This framework is then applied to a case on smart energy meters.

Chapter 25, from Inioluwa Deborah Raji, is titled "The Anatomy of AI Audits: Form, Process, and Consequences." While the previous chapter does a deep dive on transparency, Raji, in this chapter, does a deep dive on audits. Raji highlights the challenges particular to AI audits. These challenges include the context and intent of those designing the audit, the differences between internal and external stakeholders, and "in each case, the goals, perspectives and challenges faced by the auditor actively inform audit design and feed into audit processes, outcomes, and consequences." When audits are well constructed, they play a central role in evaluating and holding accountable governance actors and can play a further role in holding accountable the use of AI systems by AI Governance.

Chapter 26, from Nicol Turner-Lee, is titled "Mitigating Algorithmic Biases through Incentive-Based Rating Systems." In this contribution, after detailing the presence of bias in algorithmic decision making by governments and private companies, Turner-Lee provides a look at a particular AI Governance tool, an incentive-based rating system. Turner-Lee builds from the U.S. federal government's Energy Star program as a model to be applied to algorithmic bias. Turner-Lee argues that "rating systems can implore computer and data scientists, as well as the industries that license and disseminate algorithms, to improve their interrogation of the sociological implications, and incorporate non-technical actors and practices to inform their design and execution." It is argued that these rating systems can help mitigate biases and create further incentives for developers and companies to create AI systems that are fair, inclusive, and ethical.

The final chapter for Section V, Chapter 27, is from David Valle-Cruz and Rodrigo Sandoval-Almazán, titled "Role and Governance of Artificial Intelligence in the Public Policy Cycle." This chapter examines the public policy cycle and examine how this cycle may be influenced by Artificial Intelligence. They identify the major components of the public policy cycle including (1) agenda setting, (2) policy formulation and decision making, (3) policy implementation, and (4) policy evaluation, and then discuss the ways in which the response to the global COVID-19 pandemic highlights a variety of pathways by which AI is being further incorporated into the public policy cycle. The authors argue that the pandemic response highlights that AI may influence the policy cycle by (1) "An improved agenda setting revolutionized, faster, more efficient, with fewer data errors"; (2) "A policy formulation that contains lessons learned from previous experiences"; (3) "Faster policy implementation, closer coordination and collaboration among government agencies resulting from immediate, high-quality, and simultaneous data sharing"; and (4) "in the policy evaluation, we expect to interrelate automated learning, knowledge of computational errors, and human experience."

Section VI: AI Governance from the Ground Up

Instead of taking an expansive view of AI governance, Section VI focuses on specific stakeholders, including the public, regulators, tech workers, and the machine learning community. This section's attention to the details and nuances of various communities in the AI governance space complements the broad theory-building and policy recommendations in other chapters.

In Chapter 28 "Public Opinion Toward Artificial Intelligence," Baobao Zhang reviews existing research on the public opinion toward AI and proposes four new directions for future research. Studying public opinion toward AI is important because the public is a major stakeholder in shaping the future of the technology and should have a voice in policy discussions. Survey data worldwide show that the public is increasingly aware of AI; however, they—unlike AI researchers—tend to anthropomorphize AI. Demographic differences (including country, gender, and level of education) correlate with trust in AI in general and in specific applications such as facial recognition technology and personalization algorithms. Future research directions include studying institutional trust in actors building and deploying AI systems and investigating the relationship between attitudes and behavior.

Two of the chapters in this section focus on complexity related to the regulation of AI systems. In Chapter 29 "Adding Complexity to Advance AI Organizational Governance Models," Jasmine McNealy recommends creating networked and complex governance models to regulate AI effectively. Given the complex nature of the AI systems, traditional governance models that center power in a central agency is adequate, McNealy argues. Instead, she proposes a governance scheme involving a network of several agencies with different subject matter expertise coordinated by an administrative agency. Furthermore,

she recommends creating feedback loops of consultation and revision for regulation so that the governance scheme can be responsive.

In Chapter 30 "The Role of Workers in AI Ethics and Governance" Nataliya Nedzhvetskaya and J. S. Tan explore the role of workers in the tech industry in shaping AI governance. They expand the definition of tech workers beyond programmers and machine learning researchers to include those who label datasets and those whose work is controlled by AI systems. Tech workers play a central role in AI governance by identifying algorithmic harms as well as shaping the governance decision and responding to the decision. Analyzing over 25 collective actions that involve disputes about AI systems, the authors highlight the three types of claims that tech workers make: that they are subject to the harms of the AI system, that they should have control over the product of their labor, and that they have proximate knowledge of the AI system.

In Chapter 31, "Structured Access: An Emerging Paradigm for Safe AI Deployment," Toby Shevlane proposes that AI developers use structured capability access to share their pre-trained machine learning models to minimize the risk of misuse. While the existing norm within the AI research community is to open-source pre-trained models, doing so would risk bad actors misusing the models, including circumventing safety restrictions or reverse engineering the models. Instead, Shevlane argues that developers should allow users to access the models via cloud-based interfaces, such as application programming interfaces (APIs), to prevent users from modifying or using the models to cause harm.

In Chapter 32 "AI, Complexity, and Regulation," Laurin B. Weissinger describes the challenges of regulating AI due to the complexity of the technology and how extensive AI systems are embedded within social organizations. Furthermore, the power imbalance between those developing and deploying AI systems versus those who are impacted by these systems makes regulation even more difficult. Weissinger recommends viewing the regulation of AI through a political and economic lens rather than purely a technical one.

Section VII: Economic Dimensions of AI Governance

Markets and other economic institutions are among the most powerful forces shaping modern society. On the one hand, they have had powerful positive effects on society, generating massive increases in material prosperity that would not have been possible without market forces relentlessly steering society towards a more efficient allocation and use of resources. On the other hand, both scholars and civil society have long worried about humanity being at the mercy of blind market forces that subjugate human ethical values to market value. As a result of this tension, economic governance has been a rich and important aspect of governance more broadly.

Advances in AI may supercharge both effects. Much of the progress in AI is driven by economic forces, and AI systems—able to optimize for uni-dimensional objectives with an efficiency that we have never seen before—have the potential to greatly increase the efficiency of our resource allocation. Yet at the same time they may also perpetuate the subjugation of

human values to economic value, and to generate ever more efficient ways of undermining our human values, as observed in earlier sections of this Handbook.

This section focuses on one particular economic challenge that transformative advances in AI may pose for governance: Even though AI may lead to rapid growth in productivity and economic output, there are serious concerns about its effects on labor markets and income distribution. In particular, transformative advances in AI may put in question our current system of resource allocation, which relies heavily on distributing income to humans based on the scarcity of their labor.

The section begins with Chapter 33 by Daniel Susskind, "Technological Unemployment," which describes the economic underpinnings of how advances in technology may hurt employment and drive down wages. The chapter also provides an instructive description of how economists' thinking on the question has evolved over time. It distinguishes between two forms of unemployment: frictional technological unemployment that arises because mismatches in the labor market take time to be overcome, and structural technological unemployment that arises because overall demand for labor has declined. It then observes that one potential harbinger of the labor-saving effects of technological advances is the growing inequality experienced by many countries in recent decades, and this should make us take the concern very seriously.

Chapter 34, "Harms of AI" by Daron Acemoglu, broadens the analysis to a range of potential economic, social, and political harms of AI. The chapter identifies three main areas of concern. First, by collecting and controlling vast quantities of information, AI may pose a threat to consumer privacy, damage market competition, and manipulate consumers. Second, it argues that advances in AI may also create harm in the labor market. Building on the concerns articulated in Chapter 33, AI advancement may place an excessive focus on automating work and thereby increase inequality; it may make humans worse at judgment as machines make more decisions; and it may enable more intrusive worker monitoring. Third, advances in AI may erode our democracies through several channels: by changing our traditional means of communication, it may lead to more echo chambers and shallower political discourse; it also enables autocrats to engage in greater surveillance; and it undermines the political power of workers. The chapter ends with a passionate appeal to redirect advances in AI to steer away from the described harms.

Chapter 35 by Carles Boix, "AI and the Economic and Informational Foundations of Democracy," further analyzes the threats that AI may pose for democracy. The chapter observes that AI may lead to two fundamental transformations—greater wage inequality and higher concentration of capital ownership—that will also affect political power dynamics between workers and the owners of capital. If advances in AI lead to sufficient growth, political systems may counteract the market forces that lead towards greater inequality, but the chapter observes that large challenges remain, especially for poorer nations that experience smaller gains or even losses from AI. Moreover, AI may also increase political polarization and strengthen the hands of autocrats through more effective surveillance technologies.

In Chapter 36, "Governing AI to Advance Shared Prosperity," Katya Klinova turns to the question of how AI governance can actively contribute to redirecting AI to benefit workers rather than substituting for them. Reviewing the motivations and constraints that AI developers are subject to, the chapter identifies not only laws and market incentives, but also certain social norms, benchmarks, and visions pursued by AI developers as culprits for

an excessive focus on displacing labor. Moreover, the chapter emphasizes the importance of tools and processes that allow AI developers and policymakers to assess the impact of AI technologies on employment, wages, and job quality. It also d escribes opportunities for tangible governance interventions to steer AI towards a path of greater shared prosperity.

The last contribution to section VII is Chapter 37, "Preparing for the (Non-Existent?) Future of Work," by Anton Korinek and Megan Juelfs. The chapter starts by analyzing the technological and economic conditions that may lead to the demise of labor and squares them with the experience of the past two centuries of rising wages. Then it examines how to optimally allocate income and work in such a world. When labor demand declines sufficiently, it becomes optimal to phase out work, beginning with workers who have low labor productivity and low job satisfaction. Moreover, a basic income from capital ownership or benefits may be the only way to avoid mass misery. However, if there are positive externalities from work amenities, such as social connections or political stability, public policy should encourage work until society develops alternative ways of providing these amenities.

Section VIII: Domestic Policy Applications of AI

Section VIII examines the range of governance contexts in which AI is used domestically.

Chapter 38, entitled "Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance," presents a case study in change management for AI. The setting is controversial: the use of AI in social security claims adjudication. In this chapter, Kurt Glaze, Daniel E. Ho, Gerald K. Ray, and Christine Tsang tell the story of how, in part through individual expertise and persistence, the Social Security Administration (SSA) adopted AI systems to augment adjudication processes. The chapter analyzes what factors led to the adoption of AI and it draws lessons to inform governmental AI adoption in other contexts. Given that administrative agencies increasingly use AI to drive service delivery, this chapter is a highly relevant, timely, and instructive study.

Chapter 39, "Watching the Watchtower: A Surveillance AI Analysis and Framework" by Stephen Caines, takes a close look at surveillance technologies that are regularly employed by local governments. The chapter develops a practical framework to think about surveillance, it highlights important risks—such as function creep and mission creep—and hypothesizes the major threats and developments that AI brings to the governance horizon.

Chapter 40 by Beatriz Botero Arcila, titled "Smart City Technologies: A Political Economy Introduction to Their Governance Challenges," offers a comprehensive view on smart cities. Smart cities are recruited by two opposing narratives: one focuses exclusively on the opportunities and the other on the risks of increasing the connection and computation of urban infrastructures. With a close attention to both opportunities and risks, the chapter analyzes smart cities through the lens of political economy. This yields an analysis that is both insightful and practical—it supports scholars and policymakers to deliberate about institutional interventionsto govern this technology.

Chapter 41 is "Artificial Intelligence in Healthcare" by Nakul Aggarwal, Michael E. Matheny, Carmel Shachar, Samantha X.Y. Wang, and Sonoo Thadaney-Israni. The chapter

contextualizes the opportunities that AI brings for the diagnosis and management of patients' health within the history of the use of algorithms in healthcare. It argues that the Quintuple Aim of healthcare—patient outcomes, cost reduction, population impact, provider wellness, and equity and inclusion—still serve as a guidance to govern AI healthcare innovations. However, the authors strongly caution about the risks of algorithmic biases. Specifically, it highlights how unrepresentative datasets can exacerbate disparities and it emphasizes the need for diversity, transparency, and accountability.

"AI, Fintech, and the Evolving Regulation of Consumer Financial Privacy", Chapter 42, by Nikita Aggarwal, looks at an impactful but often overlooked domain of privacy: consumer finance. With a focus on English law, the chapter traces the legal evolution of privacy protections of consumer financial data to ideas of financial confidentiality. Consumer finance privacy law has evolved with technology from bank confidentiality to cross-sectoral data protection. The most recent developments are shaped by fintech, that is, the increasing use of AI and data-driving technologies. The chapter describes the opportunities and challenges of these developments for the regulation of consumer finance in the age of AI.

Section IX: International Politics and AI Governance

Section IX spans issues from great power competition and international politics to strategy and concerns around military applications, and to opportunities and challenges for regional governance with examples from the Global South, EU, and NATO.

One core aspect of modern governance is how nation states and transnational institutions relate to one another on the global landscape. The various ways in which these entities cooperate, compete, and behave has, over time, been greatly impacted by the available technology. This has been the case both for technological tools of communication and of destruction. Relatively new and powerful forms of instantaneous global communication have reshaped how nation states make decisions internally and how they communicate and interact with other nation states. Messages, and their various translations, can be shared almost instantaneously. At the same time, technologies of destruction have evolved to include nuclear weapons and the birth of lethal autonomous weapons. It should be clear that technological evolution has both direct and spillover effects for the ways in which nation states and transnational institutions play their games.

Advances in artificial intelligence present both opportunities and challenges for improving the coordination, decision making, and governance of nation states and transnational entities. The chapters in this section, using lenses from international relations and international politics, examine the influence of artificial intelligence for the internal decision making and corresponding behavior of individual nation states and transnational institutions and for how those institutions interact with one another.

The chapters highlight the global nature of the challenge of governing the development and use of artificial intelligence. As with the global challenges of climate change, nuclear proliferation, and war, appropriately and effectively governing the artificial intelligences that are deployed into society, and particularly those deployed in service of nation states

and transnational institutions, is a global challenge. The spread of the digital universe and the globally connected internet, the places in which artificial intelligences are deployed, is a global development that is shaping the behavior of nation states and their interactions with one another, and thus how global power and resources are distributed, amassed, and controlled. In the chapters that follow it is made abundantly clear that the socio-political ecosystem in which AI is being developed and deployed has a direct impact in shaping and being shaped by that very global political power structure. The resulting conditions and consequences are explored.

Chapter 43, from Jeffrey Ding, is titled "Dueling Perspectives in AI and US–China Relations: Technonationalism vs. Technoglobalism." Ding argues that the narrative around US–China relations relies heavily upon a frame of technonationalism which places significant emphasis on the competitive games played across nation states. Ding highlights an additional framing called technoglobalism. This frame highlights the various transnational networks already in place in which both the United States and China are heavily embedded and illustrates that the technonationalism frame is limited by not fully considering the pressures arising from these transnational networks.

Chapter 44, from Justin Key Canfil and Elsa B. Kania, is titled "Mapping State Participation in Military AI Governance Discussions." Kania and Canfil use the current debate around lethal and autonomous weapons systems (LAWS) that is under discussion at the UN Convention on Certain Conventional Weapons (CCW) as an empirical exploration to understand what types of countries are likely to engage in these conversations and why. Their empirical results reflect a nuanced picture of when and how states choose to engage in discussions on LAWS. It does seem that the disarmament agenda is supported most strongly, not necessarily by democracies, but by those in the developing world. A finding that may also reflect the disempowered role from the Global South highlighted in Chapter 48 of this section.

Chapter 45, from Michael C. Horowitz, Shira Pindyck, and Casey Mahoney, is titled "AI, the International Balance of Power, and National Security Strategy." These contributors highlight the various ways in which AI development may change both the relative power of particular nation states and the systematic way in which nation states interact, compete, and exert power. In particular, they examine how both military and non-military dimensions of state power may be influenced by AI development and how stability of the balance of power, international institutional order, and international norms may also be influenced by AI. They argue that (1) most states are in the early stages of considering how to develop and use AI, (2) states are making near term investments to hedge against first mover status by states such as the US and China, and (3) states efforts to predict and manage their security interests will continue to be complicated by the pace of private technological advance.

Chapter 46, from Charlotte Stix, is titled "The Ghost of AI Governance, Past, Present and Future: AI Governance in the European Union." In this chapter, Stix highlights the path taken by the European Union in its attempts to govern AI and contrasts this with the approach taken by the United States and by China. Stix highlights the ways in which the EU encourages ethical, trustworthy, and reliable technological development. Stix highlights the EU's focus on trustworthy AI and strengthening the AI ecosystem. In the final section, Stix points to three strategies for the EU to consider as it moves forward with AI governance including: (1) AI megaprojects, (2) AI regulatory agencies, and (3) setting standards.

Chapter 47 is co-authored by Amelia Arsenault and Sarah Kreps, and is titled "AI and International Politics." In this chapter they explore the myriad ways in which advances in

the development and deployment of AI have important corresponding consequences for the behavior of states. Kreps and Arsenault examine the general potential impacts of AI for the furtherance of authoritarianism and democracy, the global balance of power, and war. They argue that international actors may pursue AI capabilities for improving domestic governance and efficiency, capitalizing on military capabilities, power, influence, and global competition. Finally, they argue that the way in which states pursue AI capabilities, rather than radically restructuring international politics, will instead intensify the current trends. As the authors state in their conclusion, "the basic technology and algorithms behind AI simultaneously create opportunities for improved international coordination and cooperation between actors, and exacerbate risks of invasive surveillance, competition, and escalation of military tensions."

Chapter 48, from Marie-Therese Png, is titled "The Critical Roles of Global South Stakeholders in AI Governance." In this chapter Png systematically elucidates the tensions between the Global South and Global North in the framing and debates around AI governance. Png notes the general call for more inclusive AI governance and provides a pathway for generating this inclusivity across Global North and Global South individuals and institutions. Png argues that the dominant narrative of the Global North fails to internalize the concerns and challenges expressed by the Global South. Png identifies three important roles for Global South actors in AI Governance. These include (1) "as challenging functions to exclusionary governance mechanism," (2) "providing legitimate expertise in the interpretation and localization of risks-which includes a whole-systems and historic view," and (3) "providing a source of alternative governance mechanisms; e.g., South–South solidarity, co-governance, democratic accountability, and a political economy of resistance." Finally, Png also provides three proposed steps for improving AI governance processes: "(1) engage in a historical-geopolitical power analysis of structural inequality in AI governance and international legal frameworks," (2) "identify mechanisms and protocols that mitigate "paradoxes of participation" and redress institutional barriers, in order to meaningfully engage with underrepresented stakeholder groups," and (3) "co-construct and formalize roles for Global South actors to substantively engage in AI governance processes."

The final chapter of this section and for this book, Chapter 49, is from Zoe Stanley-Lockman and Lena Trabucco. It is titled "NATO's Role in Responsible AI Governance in Military Affairs." In concert with arguments from the previous chapters and the challenges AI presents to international politics, Stanley-Lockman and Trabucco highlight the role a specific international organization, NATO, can play in helping nation states navigate the emerging challenges and military risks that arise as a consequence of emerging AI capabilities and their deployment. They highlight two particular roles that NATO, given its own competencies, can deploy tools towards AI Governance: (1) strategic and policy planning, and (2) standards and certification. Stanley-Lockman and Trabucco then discuss how these two roles of NATO should be guided by the AI Governance pillars or (1) ethics and values, (2) legal norms, and (3) safety and security. They find that NATO "is an institution with considerable opportunities to shape responsible AI governance," and that "this entails urging and facilitating Allied standards and policies to establish foundations for emerging military technology built on informed and ethical principles and enhance the international security environment."

SECTION I

INTRODUCTION AND OVERVIEW

JUSTIN B. BULLOCK

CHAPTER 1

AI GOVERNANCE

Overview and Theoretical Lenses

ALLAN DAFOE

Introduction

In the coming *years*, artificial intelligence[1] will be deployed in impactful ways, such as in transportation, health, energy, news, social media, art, education, science, manufacturing, employment, surveillance, policing, and the military. As a general purpose technology (GPT) (Garfinkel, 2024; Bresnahan & Trajtenberg, 1995), the changes induced by AI will be broad, deep, and hard to foresee (Ding & Dafoe, 2021). The upsides will be substantial, but so also will be the potential disruptions and risks (Whittlestone & Clarke, 2024).

In the coming *decades*, the impacts from AI could go much further, potentially radically transforming welfare, wealth, and power to an extent greater than the nuclear revolution or industrial revolution. Speaking to this, machine learning (ML) researchers foresee the possibility of broadly human-level AI in one (eight percent) or two (22 percent) decades, and believe it more likely than not (>50 percent) by 2060.[2] The most rigorous attempt to date to forecast human-level AI based on mapping trends in hardware to estimates of the computational power of the brain reaches similar estimates (Cotra, 2020). The consequences will be profound, be they positive or negative.

The stakes are thus high that the development and deployment of AI go well. The field of AI governance seeks to understand and inform this challenge. To clarify, I will offer some definitions. *AI governance* refers (1) descriptively to the policies, norms, laws, and institutions that shape how AI is built and deployed, and (2) normatively to the aspiration that these promote good decisions (effective, safe, inclusive, legitimate, adaptive). To be clear, governance consists of much more than acts of governments; it also includes behaviors, norms, and institutions emerging from all segments of society. In one formulation, the field of AI governance studies how humanity can best navigate the transition to advanced AI systems. This chapter offers a perspective on the field, emphasizing the challenges posed by significantly more advanced AI technology.

Four Risk Clusters

To motivate this work, it can be helpful to make the potential extreme risks more concrete. Consider the following four clusters of risk, which we will discuss in more detail in the following sections.[3]

Inequality, turbulence, and authoritarianism

Declining labor share of value and a rise of winner-takes-most labor markets could erode the position of labor and the relative equality underpinning democracy (Korinek & Juelfs, 2024, Boix 2024). Digitally mediated and AI filtered communication could increase polarization, epistemic balkanization, and vulnerability to manipulation, undermining liberal societies (Acemoglu, 2024). As with prior technological revolutions, these and other shocks could destabilize the social order and give rise to radical alternatives. Totalitarianism could be made more robust by ubiquitous physical and digital surveillance, social manipulation, enhanced lie detection, and autonomous weapons.

Great-power war

Advanced AI could make crisis dynamics more complex and unpredictable, and enable faster escalation than humans could manage—a "flash war" (Scharre, 2018)—increasing the risk of inadvertent war. Advanced AI might otherwise increase the risks of war from extreme first-strike advantages, power shifts, and novel destructive capabilities (Horowitz et al., 2024).

The problems of control, alignment, and political order

AI safety is sometimes conceptualized in terms of the *control problem*, which is the problem of human intention controlling what an advanced AI does (Bostrom, 2014; Russell, 2019). An aspect of this is the *alignment problem*: constructing AI agents to have as goals the intentions of the human principal. Some experts believe controlling or aligning advanced AI systems will be difficult (Grace et al., 2018; Christian, 2020). As our AI systems increase in power, failure of control and alignment will pose ever greater risks to the users of AI and the surrounding community.

The idea of out-of-control AI systems can seem implausible to some. We can reframe this in terms of the perennial problem of political order, a central challenge of which is the alignment and control of powerful social entities such as corporations, military actors, and political parties. This political "control problem" remains unsolved in the sense that our existing solutions are patchwork and periodically fail, sometimes catastrophically, with corporate malfeasance, military coups, or unaccountable political systems (Drutman, 2020). The AI control problem can be understood as analogous to the political control problem. As AI becomes more capable, autonomous, and empowering of certain social entities, these two control problems will intertwine and compound.

Value erosion from competition

A high-stakes race (for advanced AI) can worsen outcomes by pushing parties to cut corners in safety. This structural risk (Zwetsloot & Dafoe, 2019) from competition can be generalized to any situation where there is a trade-off between anything of value and competitive advantage, and it can impact values beyond safety. Contemporary examples of values eroded from global economic competition could include sustainability, decentralized technological development, privacy, and equality. These *negative externalities* from competition can, in principle, be governed through global institutions, but adequately channeling competition can be difficult given complexity, uncertainty, rapidly evolving technology, asymmetric interests, bargaining friction, and especially great power rivalry (Coe & Vaynman, 2020; Fearon, 1995). In the long run, ungoverned military and economic competition could mean the future of humanity is pulled toward what is most adaptive within this competitive ecosystem, rather than toward what is good (for humanity, or in any other sense). Out of this anarchic competitive milieu, we might see the entrenchment and lock-in of impoverished values and forms of life (Bostrom, 2004).

Extreme risks and a holistic sensibility

Attention to the possibility of extreme and existential risks can help ensure the field invests adequately in avoiding worst case outcomes. Part of the field explicitly prioritizes attention to extreme and existential risks (as well as extreme opportunities), often theorized in terms of risks from misaligned superintelligence (Bostrom, 2014; Russell, 2019). Broadening the focus is the concept of "transformative AI" (TAI) (Gruetzemacher & Whittlestone, 2022), sometimes defined as AI which could "precipitate a transition comparable to (or more significant than) the agricultural or industrial revolution" (Karnofsky, 2016), or as AI which could lead to "radical changes in welfare, wealth, or power" (Dafoe, 2018). As the above risk clusters make clear, there are many ways that advanced AI could have extreme impacts on humanity. Analysis tends to focus on risks more than opportunities. Most believe that AI will robustly *enable* improved welfare, health, wealth, sustainability, and other social goods. Economists, for example, overwhelmingly believe AI will create benefits sufficient to make everyone better off. From this perspective, the challenge is to ensure sufficient safety and distribution of opportunity so that the benefits brought by advanced AI can be widely appreciated.

Some scholars frame different approaches as in conflict, such as between some schools of AI ethics and AI safety focused on existential risks (Piper, 2022). Such a conflictual framing is unlikely to be helpful and is often misplaced (Prunkl & Whittlestone, 2020). Often the scholarship and policy work that needs to be done to address different kinds of risk overlap considerably. We see an analogous movement in AI safety, where scholars originally prioritized thought experiments about superintelligence (Bostrom, 2014; Yudkowsky, 2008) but have increasingly built out complementary empirically informed research programs aiming for scalable advances in AI safety starting with existing systems (Amodei, et al., 2016; Hendrycks et al., 2021). Similarly, AI governance would do well to emphasize *scalable governance*: work and solutions to pressing challenges which will also be relevant to future extreme challenges. Given all this potential common interest, the field of AI governance

should be inclusive to heterogenous motivations and perspectives. A *holistic* sensibility is more likely to appreciate that the missing puzzle pieces for any particular challenge could be found scattered throughout many disciplinary domains and policy areas. Overviews such as Dafoe (2018) and this *Handbook* offer a sampling of where insights might be found. We will now turn to a theoretical framework for making sense of the impacts from AI.

Theoretical Lenses: General Purpose, Information, Intelligence

How should we think about the impacts from AI? Any theoretical framework will have to balance desiderata. We would like a framework that is at a relatively high level of abstraction, so that our insights and conceptual vocabulary generalize across issue areas; however, we also want enough structure and concreteness so that it yields rich predictions. We want a framework that is parsimonious, to be manageable; that is grounded in a compelling theoretical microfoundation and the technical features of AI; and that is close to exhaustive so as to not miss key properties.

There are many candidate properties and perspectives that we would want to highlight, such as AI as an enabling technology; the delegation of human decision-making to machines, and the encoding of politics in machines; accelerating and changing the character of decision-making processes, as well as systemic risks; accelerating economic growth, but with distributional implications; displacing labor, changing the value of capital versus labor, and impacting inequality; impacting the offense–defense balance and balance of power; and altering informational dynamics like surveillance, coordination, and human imitation.

The preceding perspectives are mostly descriptions of potential impacts from AI, but they largely do not offer microfoundation for those impacts. Instead, I will offer a framework of three theoretical lenses from which these perspectives can be derived. Each of these lenses provides microfoundations and a cognate reference class, illuminating historical analogies.

These lenses are: (1) AI as a General Purpose Technology, (2) AI as an Information Technology, and (3) AI as an Intelligence Technology. The later categories can be understood as special cases of the earlier categories (although this conceptual nesting is imperfect). While these three lenses may seem complex or high level, I believe their richness and generality sufficiently compensates. We want a theory that not only makes sense of our present intuitions, but also allows us to anticipate and make sense of the dynamics that will later emerge. The following exposition involves many theoretical claims, concisely stated so as to sketch our current best understanding of the impacts of AI; however, these propositions can and should be questioned and studied further, and thus treated as hypotheses.

AI as a general purpose technology

We can think about AI as a *general purpose technology* (GPT). A GPT can be defined as a technology that (1) provides a valuable input to many (economic and other) processes, and (2) enables important complementary innovations (for other definitions and overviews see

Garfinkel, 2024; Bresnahan, 2010; Lipsey et al., 2005). Examples include printing, steam engines, rail transport, electricity, motor vehicles, aviation, and computers. Most GPTs involve either energy production, transportation, or information processing (Lipsey et al., 2005, p. 133). GPTs are often attributed responsibility for long-run economic growth.

AI is a GPT and will plausibly be the quintessential GPT. AI can serve as a fundamental input to many processes, and is highly complementary with other processes. As Kevin Kelly (2014) put it, "Everything that we formerly electrified we will now cognitize. . . . business plans of the next 10,000 startups are easy to forecast: *Take X and add AI*." GPTs tend to be more transformative the more they are "capable of ongoing [substantial] technical improvement" (Bresnahan, 2010), which seems to be true of AI: we are still in early days of AI development, and the ceiling of potential capability likely exceeds human-level. Finally, given the plausible pace of developments, political-economic transformations from AI are likely to come more quickly than they have from most previous GPTs.

GPTs tend to have a set of important properties, which AI will likely also possess. First, GPTs grow the economy, often radically so; in fact, the concept of GPT was largely conceived to explain growth in "total factor productivity," which is a crucial component of long run economic growth. We can conceptualize this growth as arising from increases in *efficiency*, where the GPT reduces the costs of inputs to existing processes, and from *enabling* new processes altogether. The potentially Pareto-improving character of GPTs is true of AI: in principle, if deployed well and if losers are compensated, AI presents a profoundly positive opportunity for all people and groups to advance their interests, to a magnitude comparable to the industrial revolution.

However (and second), GPTs tend to be *disruptive* of existing processes, and thus also disruptive of social-political relations that depend on those processes. They tend to have substantial *distributional consequences*: shifting power and wealth, providing opportunities for certain groups, companies, and countries to rise and fall. They impose (short-run) *displacement costs* on certain groups and economic factors (e.g., land, certain kinds of capital); these costs are often not easy to identify, making it hard to insure against them or contract over them (Korinek & Stiglitz, 2019). Although the earliest versions of GPTs may appear harmless and of little utility—a cumbersome printing press; a slow prototype railway; a massive hard-to-program computer—after several generations of improvement, deployment, complementary innovations, and adaptation, their cumulative impact can be revolutionary. Even while aggregate wealth increases, some individuals, groups, countries, ideologies, and cultures will lose from these changes, if only positional goods like status. Although the net impact of the past two centuries has been favorable to labor and liberal institutions, this arguably depended on the extent to which labor and liberal institutions were (economic and military) *complements* to the new technological ecosystem, which may not continue indefinitely.

Third, anticipation of disruption can *mobilize potential losers*, and cause social conflict. Workers, firms, and asset holders who fear being displaced may resist the technology, or seek political protections; the effects of this resistance range from minor regulatory protections to revolutions (Frey, 2019). At the international level, the (anticipated) rise and fall of countries, and the scramble for new strategic resources and capabilities, can precipitate aggressive actions and war (Horowitz et al., 2018).

Fourth, many GPTs are *strategic*, in the sense of being *essential to the military-industrial base and national power*; AI is one such strategic GPT (Ding & Dafoe, 2021). Those groups

and countries that successfully harness strategic GPTs gain in relative wealth and power; in fact, possession and deployment of their mature variants is close to a necessary condition of great power status. They thus become a site of strategic investment and rivalry. GPTs often give rise to critical military assets and are thus of interest to militaries.

Relatedly, GPTs are *dual-use:* they have both peaceful beneficial applications, and dangerous/military applications, and it is often difficult to separate these. They are sometimes developed in the military sector, sometimes the civilian, but have implications for both. For technologies with this inseparable dual-use character, arms control is especially difficult.

Each of these implications will apply to AI. By recognizing that these implications are not novel to AI, but are shared by other GPTs, we can learn from historical experience with this broader reference class.

AI as an information technology

A second theoretical lens regards AI as an *information technology*. Information is critical: for economic production, for coordination and identity, for power and bargaining, and for democratic oversight and authoritarian repression. Historical information technologies have had profound impacts in generalizable ways.

An information technology is one that improves the production, compression, transmission, reproduction, enhancement, storage, control, or use of information. AI will enhance the technical possibilities for each of these processes, which will then complement the others. For example, human–machine communication will be improved through natural language understanding, bidirectional oral communication, interpretable gestures, affect and psychological inference, and contextual understanding; this will then improve the production, enhancement, and expression of information. AI assists in the *compression* of large datasets into smaller generically usable datasets, such as when converting streams of video of a pedestrian square into a digital record of who was where, when, and doing what. AI will *enhance* data by making it more searchable and readable, and by identifying useful features, which may span modalities. And of course, AI will make possible a massive amount of new uses for information, on the order of trillions of dollars' worth. AI will thus be an information technology, and it will amplify other information technologies.

Some information technologies have been GPTs, inheriting the properties of the GPTs discussed above. For example, speech and culture, writing, and the printing press were crucial for the rise of, respectively, homo-sapiens, civilization, and the nation-state; the telegraph and radio enabled extensive knock-on innovations, transforming war, commerce, and political order. However, information technologies have additional distinctive properties, especially if we focus on the most recent trends in digitization and digital services.

Economic implications: Increasing returns and distribution

Information technologies, and especially digital services, tend to have substantial *economies of scale*. This arises foremost because these processes involve low *marginal costs* (e.g., reproduction and transmission of information), relative to the *fixed costs* (e.g., production of a movie). A firm makes a massive ("fixed") investment to develop mapping and navigation

services, and then pays a negligible cost for providing that service to the marginal user. To make this concrete for ML, the compute costs of inventing and training a new model are often orders of magnitude more than the later costs of deploying an instance of it. A second dynamic are *network economies* arising from (access restricted) communication networks: a language, telegraph network, phone network, operating system, and social network is more valuable the more other users are on it, leading to high returns to scale. These two features tend to concentrate the *global market structure* of information industries, favoring one or a few networks or firms (more on this below). A related implication is that information technologies tend to produce *winner-takes-most* labor markets, where a few superstar actors, writers, athletes, researchers, designers, entrepreneurs, and CEOs can capture most of the value in their market (Jones & Tonetti, 2020).

The preceding dynamics push toward greater *income inequality* (Korinek & Stiglitz, 2019), to firms, individuals, and even possibly countries. However, information technology has a strong countervailing valence toward *consumption equality* because *information wants to be free*, being non-rival and hard-to-exclude. (1) Information is hard to hold on to. (i) Sometimes just the knowledge that something can be done, or the broad contours of how it is done, is a sufficient clue to dramatically accelerate a competitor's R&D catch-up. (ii) It is difficult to provide many information services without the recipient being able to copy and reproduce it, hence the elaborate (and porous) legal and hardware protections for intellectual property. (iii) The direct costs of intellectual property theft, to the thief, are often not prohibitive (as compared with other kinds of theft, such as natural resource theft); if an employee is willing to disclose information, business secrets and digital files can often be exfiltrated. (2) Ignoring the need to fund innovation, the socially efficient arrangement is to provide goods and services at their marginal cost, which, in the digital realm, is often close to zero. This arrangement can be achieved through public interest services (e.g., the openness norm in scientific publishing; services like Wikipedia), through limits on intellectual property (e.g., copyright limits, which enables services like Project Gutenberg), and through market competition that leads to inexpensive services (exemplified by the many free or ad-based digital services). It is hard to estimate, but plausibly the value today of free services to individuals with a smartphone is worth tens of thousands of dollars per year per person (Brynjolfsson et al., 2019). Thus, while information technologies may imbalance the income distribution, it could balance the distribution of consumer welfare. Consider a billionaire: the books, movies, video games, navigation apps, and social media services they use are largely accessible to the median wage earner.

Coordination and identity

Information technologies facilitate communication and coordination, but the political impacts are often ambiguous: innovations may strengthen or undermine existing communities and power centers.

First, the economies of scale of information technologies, and complementary adaptation like standardization (e.g., in language, typography, style guides, ICANN), encourage *broader collective identities*, as information consumption can shift away from the former monopoly of local sources. This dynamic is present in the creation of national identity from more disparate local identities (Anderson, 1991), and has fueled and continues to

fuel cosmopolitanism and liberalism through literature, global news, Hollywood, and the internet.

On the other hand, by allowing spatially distributed individuals with common interests to better communicate and coordinate, information technologies may support *narrow spatially distributed* identities, whose interests may be contrary to incumbents. Examples include global ideologies and movements (e.g., communism, environmentalism, Al-Qaeda), religions (e.g., Protestantism), and other cultural identities. Some information technologies have thus been critical in undermining existing power centers through cultural revolutions and the rise of complex spatially distributed communities. A similar proliferation of smaller organizational forms has taken place in the economy, with the information technology enabled rise of boutique firms and the gig economy.

Many of the potential impacts of AI can be interpreted through this lens of how it will structure the coordination of political communities, such as in discussions of epistemic security and the political valence of AI.

Power

Information technologies can shift power within a relationship, such as by making it easier (or harder) for one party to monitor the other or monopolize critical information. In situations of imperfect information, such as bargaining situations or principal–agent relations, becoming more informed is often critical for the distribution of value: it provides *information rents*. Information is often critical in adversarial contests, as it may help identify the plans, and physical and political vulnerabilities, of adversaries. Offering a rough proxy for the importance of information for international hard power, the U.S. intelligence budget is 10 percent of its total military budget. In coup attempts, be they of the state or boardroom, "information is the greatest asset" (Luttwak, 1968, p. 82), with attempts often succeeding or failing depending on the timing of when the incumbent learns about the attempt. Information is critical for domestic governance, be it effective democratic oversight or totalitarian suppression. Information technology is transforming privacy, plausibly weakening individuals' privacy against authorities, but strengthening it against social peers (Garfinkel, 2020).

The centralization of control depends on the ability of the authority to adequately monitor and communicate with its agents. The telegraph and radio dramatically curtailed the autonomy of ambassadors and ship captains. Remote and autonomous weapons will similarly empower commanders to execute orders without delegating through officers (who might object, for example, to orders to shoot civilians). Technological trends are not always toward greater centralized control, however, as exemplified by the printing press and the invention of RSA (Rivest-Shamir-Adelman) and Pretty Good Privacy encryption.

Information technologies generally increase the (economic and military) value of information and its infrastructure. This is evident in the rise of military activities in cyberspace and information operations via social media, and we can expect this trend to continue.

Information can move at much faster speeds than other processes: from chains of smoke signals in ancient China traveling hundreds of kilometers an hour, to the (apocryphal) use of carrier pigeons by the Rothchild's to learn of the outcome of Waterloo before others, to

contemporary traders investing billions to construct inter-exchange fiber-optics and microwave beams for advantages of milliseconds. This acceleration from information-based dynamics can lead to an acceleration of crises, as exemplified by financial "Flash Crashes" where trillions of dollars in value can disappear in minutes.

The net effect of any information technology on politics and power is often hard to know in advance. It remains too early to say with confidence whether AI will strengthen the state, weaken it, or lead it to be subsumed or transformed. But it is clear that information is a critical resource for political dynamics, and AI will amplify the value and impact of that resource. These themes will recur below.

AI as an intelligence technology

The third theoretical lens involves understanding AI as a technology of *intelligence*: an innovation in the ability (of some entity) to solve cognitive tasks (Hernández-Orallo, 2017). Of the three perspectives on AI proposed here, this third is the least well-developed in the literature; however, it arguably illuminates the most important impacts of artificial intelligence. The advent of AI thus demands that we make sense of the broader reference class of intelligence and intelligence technologies. (I will only use the term technology occasionally for this lens, because it is an imperfect fit for some kinds of intelligence innovations, like the use of humans as advisors.)

Intelligence technologies can vary on a number of dimensions. One important distinction is between *tools*, on one end of the spectrum, and *systems* and *agents* on the other. Examples of *tools* include an abacus, a dictionary, and a notepad. These are *narrow*—they are designed to perform some specific function, and they are not meant to impact the world beyond that narrow use. These are *not autonomous*—they require a user to have impact in the world, by integrating them into some broader goal-directed process. Other intelligence technologies are more general and autonomous; these, which are often the most impactful, we can call *systems* or *agents*, with "agents" denoting those that behave more like coherent goal-seeking entities. Examples of (intelligence augmenting) *systems* include the price mechanism in a free market, language, bureaucracy, peer review in science, and evolved institutions, like the justice system and law. Examples of *agents* are chief advisors to a monarch (e.g., the Grand Vizier), the general staff for the military, a corporation, or a deeply socialized bureaucracy. List and Pettit (2011) examine the concept of agency applied to groups. Danzig (2022) similarly analogizes AI to bureaucracies and markets and considers with each the alignment and control problems.

We can draw out several high-level properties of intelligence technologies, echoing implications we saw with general purpose technologies and information technologies. They are often critical for military and economic survival; consider the military general staff or the use of the market to allocate resources. They often transform the character of the largest political entities. Human tribes, the Neolithic state, the medieval state, and the modern state each arose in part from improvements in intelligence technologies.

Historically, intelligence technologies both substitute for and complement (other) human cognitive labor. The use of machine calculators substituted for human "calculators," but also complemented (made more valuable) other mathematical skills. The rise of a

competent bureaucracy substitutes for the actions of an individual minister, who may have formerly monopolized this aspect of policy, but also can complement a decision-maker with good judgment. An important question concerns the extent to which future AI will complement, or substitute, for human cognitive labor (Brynjolfsson, 2022), as this could have profound implications for labor share of value, inequality, and growth rates (as capital can grow itself).

Bias, alignment, and control

Perhaps, most importantly, intelligence entities often pose challenges of bias, alignment, and control.

Even simple tools can *bias* decision-making: leading us to pursue more of that which fits the tool or is made salient by the tool: *to a person with a hammer, everything looks like a nail*. Arguably, early states were biased toward legible social arrangements (Scott, 2008), contemporary policymakers focus too much on GDP (rather than actual wealth and wellbeing) (Sen et al., 2010), and social media companies optimized excessively for metrics like engagement. Because there are often political implications to any decision-making process, the introduction of cognitive tools which shape those processes themselves have political implications. We are seeing this politics of cognitive tools in the use of AI for decision-making related to employment, crime and justice, and social relationships, but also in how tools like email and Twitter shape how people communicate, deliberate, and work.

More problematically, systems and agents may not be *aligned* with or under the *control* of the principal, such that increasing the power of the system/agent will systematically lead to outcomes in conflict with the interests of the principal. An example of system misalignment is when the market loses the ability to allocate resources well when there are significant externalities. Corporate scandals offer examples of agent misalignment, such as the Enron scandal in which Enron used complex accounting practices, and then colluded with the Andersen accounting firm to misrepresent financial performance. Civilian control over the military offers another prominent category of periodic agent misalignment, exemplified by the Kennedy administration's struggles with the Joint Chiefs during the Cuban Missile Crisis (Allison, 1969), or the Obama administration's struggles with U.S. military leaders over troop requests for Afghanistan (Obama, 2020).

These problems of alignment and control can be understood as a form of Principal-Agent problem, where the agent's advantage over the principal is not just one of information but also of potentially vastly superior intelligence (Young et al., 2019). The principal may not even know what questions to ask, where to look, or have the concepts to make sense of the problem. Solutions to this problem have been explored throughout the social sciences, and in work on national governance and corporate governance, and include mechanisms for oversight, transparency, whistleblowers, representation, and other aspects of institution design. Presently, work on the problem of alignment and control for AI is almost exclusively being done by AI researchers, but given this consilience the work would benefit from experts in social science and governance (Irving & Askell, 2019).

In conclusion, it is this third lens of *intelligence* which makes clear the full extent to which AI will be transformative. Our social order depends on the alignment and control of (human and organizational) intelligence; as we augment social entities with machine

intelligence, problems of alignment and control will become ever more complex and critical.

Governance and Anarchy

Institutional fit and externalities

Governance involves shaping behavior to achieve social goals through institutions. "Institutions" are understood to refer to the full spectrum of social structures which shape behavior, including norms, rituals, rules, organizations, regulations, regulatory bodies, and legislatures (North, 1991).

A particularly useful conceptual tool (from economics) is that of *externalities*, which refers to byproduct impacts of an individual or group's actions on others, be they positive or negative. The central insight is that there are opportunities for institutions to increase overall welfare by discouraging behavior with negative externalities (e.g., pollution) and encouraging behavior with positive externalities (e.g., innovations); institutions can be used to "internalize" or manage externalities.

We can then conceptualize a governance issue by examining the kinds of externalities involved—what kinds of social dilemmas emerge—to see what stakeholders, interests, and mechanisms need to be included in any institution to address them. Such a functionalist approach to institutions is common in the discipline of economics, and in rationalist approaches in political and policy sciences (e.g., Koremenos et al., 2001).

Thus, when confronting a problem of governance, we can start by asking what properties the institution will need to adequately shape behavior toward the intended social goals. What are the externalities that need to be internalized, and over what political spaces do they span? Do existing or hypothesized institutions have the needed:

- spatial remit?
- issue area remit?
- political remit, in the sense that they adequately and legitimately represent the relevant stakeholders?
- technical competence?
- institutional competence?
- influence, such as ability to sufficiently shape material incentives?

To exemplify this functionalist approach, let us consider the relatively unpoliticized issue of the governance of self-driving vehicles. The primary interests here are the safety of citizens, mobility of travelers, and business interests of vehicle producers; the primary externality is the effect of driving algorithms on other road-users. Existing traffic safety agencies have a default presumption of institutional fit for regulating this domain, given the apparent similarities to their existing remit, but we can then ask specifically how regulating self-driving vehicles might differ from regulating human-driven vehicles. We will need new methods for evaluating safety, including leveraging the opportunities from fleet crash statistics, and from evaluating algorithm performance in simulators or test environments.

We have opportunities to recommend (or require) new forms of best practice, such as related to privacy, driver attention, and the storing and sharing of data from crashes. There will be implications for the pricing of insurance depending on the sharing of sensor data (when the human was driving) or the user's choice to let the car drive. Legal institutions will need to learn how best to attribute liability between producers and users. We need to manage new risks, such as from vulnerability to hacking, and new scales of risk, such as from the possibility of a whole fleet being hacked. As an emerging industry, there may be new considerations related to supporting innovation, and thus new stakeholders and new needed technical competencies. There will be new coordination opportunities related to the building of cooperative algorithms (Dafoe et al., 2020; Dafoe et al., 2021), the setting of standards between companies, and for the creation of smart infrastructure. There will likely be trade benefits to harmonizing regulations across borders, as well international tensions around strategic barriers to trade. Some issues will grab the attention of publics and elites out of proportion to their policy importance; for example, the question of how algorithms should ethically resolve variants of the Trolley problem* is philosophically engaging and unsettling, but largely irrelevant to the work needed to improve human well-being. Having done such an analysis of the new governance challenges, we are then in a better position to diagnose the kinds of institutions we are likely to need to well govern the domain.

The greater the distance an emerging governance area is from an existing legitimate competent institution, the greater challenges we will likely face in adapting or building adequate institutions. Some issues, although they may involve clear social benefits, can still fail if the needed institutional adaptation is just too great. As an example, consider the benefits of having every out-of-copyright book digitally available for free at local libraries; this profound public good is not being provided, not for want of a party willing to scan and provide this service, but because of congressional inability to update copyright law to legalize it (and the Department of Justice's discomfort with permitting a settlement of a class action lawsuit that would have legalized this, but only for one firm) (Somers, 2017).

Institutional adaptation or innovation will be especially difficult when the issues fall in, or are framed as part of, unresolved social conflicts. Here we may lack an overarching consensus on social goals, and we may lack legitimate institutions with which to work. Further, by touching on these sites of conflict, the issue may itself become a battleground for the conflict, making engagement less about the issue than about the broader conflict. We could call these *deeply politicized* governance issues, understood as issues which connect to significant political conflicts at the highest levels of effective political order (i.e., usually the country). Specifically, I will reflect on the difficulty of AI governance in the presence of domestic political conflicts—between domestic political groups, and between authorities and citizens—and great power security competition.

* The "Trolley problem" is a set of hypothetical ethical dilemmas about sacrificing one person to save many. It reveals tensions between utilitarian and various forms of deontological and other guides to ethical behavior. With self-driving cars it received notable attention from the "Moral Machines experiment" (https://www.nature.com/articles/s41586-018-0637-6) which provcatively asked subjects when a self-driving car, forced to make a choice, should kill its passengers vs a set of pedestrirans (of varying demographic profile). Such choices are exceedingly rare.

Domestic conflicts

Many countries have significant social conflicts. For example, within the United States there is the left–right political cleavage, sometimes also referenced by the terms "culture war" or "polarization." A salient example in AI governance is social media moderation, where conservatives and liberals, with different emphases, worry about censorship, bias against one's views, proliferation of "fake news," filter bubbles, foreign intelligence operations, and the mobilization of hostile social movements. Society here seems to fundamentally disagree about what it would mean to have effective, safe, legitimate social media moderation. Given deep distrust across groups about whether they have compatible social goals, it will be hard to build institutions that are widely regarded as achieving their social goals.

Another example is fairness in the use of classifiers in sensitive domains such as criminal justice, loan decisions, education, or employment. ProPublica famously sparked a discussion about racial bias in the use of algorithms for predicting the likelihood of individuals' future criminal activity, as sometimes used in parole appeal hearings, or bail or sentencing decisions (Angwin et al., 2016). ProPublica investigated an algorithmic system commonly used in the United States and reported that the false positive rate was significantly lower for white defendants than for Black defendants (i.e., when certain kinds of errors were made by the algorithm, those errors tended more often to help whites and hurt Blacks). Later research clarified that if any demographic differences in the three key quantities of false positive rates, false negative rates, or calibration are interpreted as evidence of "bias," then (if the classifier does not perfectly predict behavior and true crime rates differ according to the sensitive demographic trait) "bias is mathematically inevitable." By this definition, *every classifier* is "biased"; we need a more refined understanding of fairness to guide algorithmic governance.

More recent research has examined classifiers which optimize for different weightings of these quantities (Hardt et al., 2016), or indeed other relevant quantities such as conditional rates, occasionally identifying opportunities for strict improvements (Zafar et al., 2017). This case illustrates how a new capability—algorithmic decision-making—can force us to be more precise and explicit about what exactly we mean by bias, thereby opening a political debate over issues for which we lack thorough consensus, principles, and institutions (Kroll et al., 2017; Coyle & Weller, 2020). In short, one reason algorithmic bias is such a difficult governance issue is because, as a political community, we have not yet reached sufficient agreement about the principles for decision-making in sensitive domains.

AI does not just occasionally touch on pre-existing social conflicts. As smart sensors record ever more human behavior, and as decisions move from the black box of the human brain into manipulable and auditable algorithms, *AI will systematically expand the terrain for subjecting decisions to political control*. The direct effects of this can be good or bad, depending on whether political control over particular decisions would be good or bad. A systematic effect, though, is that it *increases the stakes of political contestation*. AI could thus inflame domestic political conflicts in a way analogous to how the discovery of natural resources could inflame a disputed border. We may come to regard the pre-AI era as one of significant autonomy for individuals—be they citizens, workers, or students; police officers, managers, or educators.

This systematic expansion of the possibility of political control is especially salient in conflicts between state authorities and citizens, such as over the appropriate extent of state surveillance. The existing social contract may not have been thought through and institutionalized for these new domains for state authority. Consider how Edward Snowden's leaks revealed secret U.S. surveillance activities which had not been thoroughly publicly debated. Following the leaks, Congress did legislate and clarify the institutions around domestic surveillance. In general, when authorities and citizens in a country are at an uneasy status quo, the expansion of opportunities for political control brought by advances in AI are likely to shift the balance of power toward the state. This dynamic is seen in work on "digital authoritarianism" (Polyakova & Meserole, 2019); more work is needed to build a vigorous positive agenda of how advances in AI can strengthen liberal institutions (Hwang, 2020).

Great power security competition

The most recalcitrant and encompassing conflict is that between rival great powers—those states with the ability to exert influence on a global scale. International relations scholars characterize the enduring structural condition facing great powers as that of *anarchy*: there is no higher authority who can make and enforce laws, and so everyone's final recourse is force. Under anarchy, the shadow of power darkens all diplomacy; everything can be renegotiated, threatened, and destroyed. No one can rest secure. This vulnerability can then produce a security dilemma in which each state seeks security through their military, but in so doing makes others feel more vulnerable. The costs of this anarchy are significant: the world spends $2 trillion per year on the military (Stockholm International Peace Research Institute, n.d.), and lives with an ongoing risk of catastrophe and nuclear holocaust from thousands of nuclear warheads, 2,000 of which remain on high alert (Kristensen et al., 2022). Inadequate solutions for other global public goods—such as climate change, global trade, and pandemic preparedness—are also plausibly consequences of global anarchy. To be clear, there are also significant risks from any attempted political remedy to global anarchy, namely from excessive political centralization (and, put strongly, global totalitarianism; Caplan, 2008).

AI governance, therefore, will be especially challenging for those issues that are deeply connected to great power security competition. The needed institutions for these are often global, but we may lack sufficient consensus about our social goals at that scale. Even when we have consensus (e.g., nuclear war is bad), the bargaining and security dynamics induced by anarchy may mean that we still cannot build the needed institutions to achieve our goals (e.g., our inability to get the world's nuclear arsenal into the small hundreds, let alone to zero).

A first cluster of governance issues in this space concerns the development of AI for lethal autonomous weapons (LAWs), cyber operations, and foreign influence operations. For each of these, but especially LAWs,[4] it is often argued that it would be desirable for countries to restrain their development and deployment of AI in certain respects. Clarifying and reinforcing norms around desirable development can shape military behavior; consider how the nuclear taboo has held back the use of nuclear weapons since 1945 (Tannenwald, 1999). However, the logic of security competition relentlessly bears down. The U.S. Department of Defense for a long-time articulated a policy that prioritized maintaining a

human-in-the-loop. However, "when instant response is imperative, even [the U.S.] Defense Department's proponents of humans in the loop concede that their desired human control cannot be achieved" (Danzig, 2018). LaPointe and Levin (2016) conclude their article on LAWs by stating: "Military superpowers in the next century will have superior autonomous capabilities, or they will not be superpowers."

A second cluster of issues arises from the decoupling of the supply chains, commerce, and research between China and the West. China has long imposed significant constraints on Western tech companies, and it has now effectively banned most Western AI services. During the past years, the United States has escalated its efforts to gain independence in its supply chain for chips (such as through the CHIPS for America Act, which is likely to provide approximately $50 billion [Arcuri, 2022]), and ensure dependence in China's (such as by blocking export of extreme ultraviolet lithography technology from the Netherlands' ASML). Other areas of decoupling are in ML research—collaborations with Chinese groups are increasingly politically scrutinized—and AI policy—consider that China is notably not included in the Global Partnership on AI, one of the most significant international initiatives on AI (Bengio & Chatila, 2021).

The AI race

An often invoked metaphor is that we are in, or entering, an *AI arms race* (Zwetsloot et al., 2018; Scharre, 2021). This metaphor is usually meant to communicate that (1) investments in AI are increasing rapidly (2) because of a perception of a contest for very large (geopolitical) stakes, (3) and these investments are militarily relevant. This metaphor is clearly sometimes abused, such as when it is invoked to reference the "arms race" in the cost of tech talent. At present, the "arms" modifier is largely literally off-point because most of the geopolitical activity in AI is not about weapons per se, but is instead about supply chains, infrastructure, industrial base, strategic industries, scientific capability, and prestige achievements. What is not in doubt is that great powers perceive leadership in AI to be critical for future wealth and power. We might more accurately call this strategic technology competition or, more abstractly, *the AI race*.

A related metaphor, and set of ideas, is that of the race to the bottom or "race to the precipice" (Armstrong et al., 2016; Askell et al., 2019). This metaphor emphasizes how race settings—where actors perceive large gains from relative advantage—can induce actors to "cut corners," exposing the world to risks that they would otherwise prudently avoid. There are many examples of commercial races where pressure to generate profit appear to have led firms to generate excessive risks. Recent fatal AI related examples include Boeing's approach to its 737 MAX, and Uber's efforts to catch up in self-driving (although see Hunt [2020] for a positive appraisal of the strength of aviation safety institutions).

A worry is that a geopolitical race to the bottom could take place. Great powers could come to perceive themselves in a strategically decisive race for powerful AI, analogous to nuclear technology in 1945, and rush into hurried crash programs, "AI Manhattan Projects." Especially if occurring in a period of geopolitical tension, as arms races tend to, each military may be tempted to deploy powerful, but not entirely reliable, AI systems in cyber and kinetic conflict. To make the accident risks concrete, consider that in cyber-war there are likely to be significant advantages to speed, such that a human-in-the-loop would be untenable. There would likely be a possibility of unintended behavior from ML systems,

especially in complex adversarial settings. Finally, there may be incentives to deploy AI cyber systems at scale, so any unexpected extreme behavior may have broad impacts. One current aspiration is to insulate nuclear command and control from ML powered cyber-operations so as to limit the destruction from an AI cyber accident.

How risky should we expect such a race to be? A useful starting model is the two-player strategic game known as the war of attrition, where players "bid" up the risk level or costs of conflict, until one player concedes (this is equivalent to an all pay auction, but where the "auction revenue" are the risks or costs of conflict). This is similar to the game of chicken, but with a continuous action space, and is a canonical model of actual wars of attrition and nuclear brinkmanship. Given typical simplifying assumptions such that players are rational and have common knowledge of the game, a typical result is that such two player auctions will "generate expected revenue" from each player of one-third of the expected value of the prize to each participant (Nisan et al., 2007, p. 236). To make this concrete, suppose that "the prize" is perceived to the decision maker as of "existential stakes," and so similar value, relative to the status quo, as the status quo is to nuclear war; then that decision maker should be willing to race up to a 33 percent chance of nuclear war. A related modeling literature (Nitzan, 1994) finds that 50 percent of the rents are dissipated in two player rent-seeking contests. So, the glass is half full? The bad news is that if a prize is sufficiently attractive, such as might be perceived around a geopolitically decisive technological advantage, decision makers may be rationally willing to expose the world to a significant risk of devastation. The good news, in this model, is that these rational actors don't race *all the way* to the bottom.

However, a real-world geopolitical race may be much worse. Decision makers may not have "common knowledge" of the game, but instead may be strategically encountering it for the first time. Economists joke that if you want to quickly raise $100 dollars, host an all-pay auction for $10; invariably some participants will fail to realize that the best strategy for all but one person is to not play, and these participants will fall into an escalation spiral committing ever more funds. Further, the risks we are contemplating are novel "tail risks," whose novelty and distance from experience will make them hard to reliably forecast. In addition, decision makers may not be rational in various senses assumed by the model: they may be overconfident, or place intrinsic value on relative outcomes or winning. When the leader of a proud country perceives honor to be at stake in a crisis, or regime survival, the costs of backing down can become much greater to them than the material issues at stake.

The above models assume that leaders can accurately perceive each other's risky behavior. If instead risky behavior lacks a publicly observable signal or is only observed with significant noise, then it will be even harder to build viable norms and institutions for mutual restraint. The above models assume that leaders have a shared understanding of AI risk. What if, instead, AI safety is highly theory-dependent? Through a winner's curse dynamic, and psychological and organizational rationalization, individuals and organizations may come to systematically perceive their own behavior as safer than that of others. This can lead to an escalation spiral, where each perceives that the other is behaving much more recklessly than they are, and in turn escalates their risk taking. There are no doubt many other psychological, organizational, and political pathologies that could further exacerbate the risks from geopolitical crises.

Escaping race dynamics

How do we escape such a dangerous race dynamic? At one end of the solution space are unilateral steps to deescalate. The above models are all based on such unilateral solutions when all racers (but one) stop racing because they perceive the risks to be excessive. Interventions that reduce the chances of underestimating risks, or that make it easier for leaders to opt of the race, therefore should be risk-reducing; however, as is often the case in models of coercion, this may also induce the other party to increase their ambition or aggression (the net effect is still usually to reduce overall risks, just by less than the direct effect, see Banks, 1990; Polachek & Xiang, 2010). Other "unilateral" solutions involve the use of force to compel an end to the race, although the possibility and execution of such options can be as risky as the race itself.

At the other end are cooperative steps to achieve *mutual restraint* (Barrett, 2007) to make the race less risky or intense. Here we seek to construct norms, treaties, and institutions to change "the rules of the game," so that racers internalize the risks. These solutions typically require actors to reach sufficient agreement about what actions are unacceptably risky, devise means to observe compliance, and identify sufficient incentives (usually sanctions) to induce compliance. A core strategic obstacle is the transparency-security tradeoff: the arrangement must provide sufficient transparency about arming behavior to assure the other party, while minimizing the kind of transparency that compromises the security of the monitored party (Coe & Vaynman, 2020). More generally, lessons from arms control of strategic technologies, such as nuclear weapons, can be instructive (Maas, 2019; Zaidi & Dafoe, 2021; Scharre, 2021).

Third-party organizations and global institutions may be crucial in reducing the risks from an AI race by moving key functions of an agreement out of the halls of diplomacy into (ideally) an impartial specialized organization devoted to the function. Trusted third parties—such as safety-focused organizations—can help articulate focal norms and standards of safe conduct. To the extent the AI race is driven by prestige motivations, as was probably true for the Space Race, third parties may be able to channel the prestige gradient toward more prosocial endeavors (Barnhart, 2022). Third-party institutions may be invaluable for verifying and ruling on non-compliance, as the WTO and IAEA do in their respective areas; third-party institutions can also help overcome disclosure dilemmas, enabling better sharing of information (Carnegie & Carson, 2020). Although far from our current geopolitical realities, third-party institutions may even take on forms of hard power, such as imposing sanctions or directly controlling materials and activities (as was proposed for the Atomic Development Authority).

In contrast to nuclear weapons, however, military AI applications are likely to be even more difficult to control. Zaidi and Dafoe (2021) summarize:

1. **Dual use**: Powerful and dangerous AI applications will likely be harder to separate from beneficial applications, relative to nuclear technology.
2. **Value:** The economic, scientific, and other non-military value from general advances in AI will greatly exceed that from nuclear technology.
3. **Diffusion**: AI assets and innovation are much more diffused globally.
4. **Discernibility of risks**: The risks from nuclear weapons are likely easier to understand.

5. **Verification and control**: It is easier to unilaterally verify nuclear developments (nuclear tests, ICBM deployments) than deployments of dangerous cyber-weapons, and it appears easier to control key chokepoints in the development of nuclear weapons, such as with nuclear fuel and centrifuge technology. For applications of AI to cyber operations, computer hardware would be among the most tangible components, but remains deeply dual-use.
6. **Strategic gradient**: To a first approximation, the strategic value of development and innovation in nuclear weapons plateaued once a state had secure second-strike capability with thermonuclear weapons. The marginal value of the 300th warhead is small (which is why China has retained an arsenal less than 300 for nearly six decades). Decades of innovation largely haven't destabilized mutual assured destruction between the nuclear powers. AI may have a persistently steep strategic gradient, incentivizing more racing and increasing the volatility in power.

Value erosion

The discussion here, and in the literature, largely focuses on risks of catastrophic accidents. However, advanced AI and great power security competition can mix to bad effect in more subtle, gradual ways. Most of this chapter's concerns can be theorized as arising because of a safety-performance tradeoff, and the competitive incentives that push actors to trade away safety for competitive performance. We can generalize this mechanism: any time there is a tradeoff between something of value and performance in a high stakes contest, competitive pressures can push decision makers to sacrifice that value. Contemporary examples of values being eroded by global economic competition could include optimally competitive markets, privacy, and relative equality. Mark Zuckerberg captured this logic in his prepared talking points for Congress: in response to the idea that Facebook should be broken up, Zuckerberg intended to respond that doing so would undermine a "key asset for America" and "strengthen Chinese companies" (Foroohar, 2019). In the long run, competitive dynamics could lead to the proliferation of systems (organizational types, countries, or autonomous AIs) which lock-in undesirable values. I refer to this dynamic as *value erosion*; Nick Bostrom (2004) discusses this in "The Future of Human Evolution"; Paul Christiano (2019) has referred to the rise of "greedy patterns"; Robin Hanson's (2016) *Age of Em* scenario involves loss of most value that is not adapted to ongoing AI market competition.

A common objection to the idea of value erosion is that history has seen long-term trends favoring humanity, and so empirically it does not seem like technological advances and military-economic competition lead to this kind of malign evolution. This perspective is usually coupled with a view of history as driven by the agency and intentionality of key decision makers, such as the Founding Fathers of the United States. This perspective may fail to appreciate how circumstances for humans did not obviously improve following previous technological revolutions, such as the neolithic revolution (Karnofsky, 2021). Further, the increase in human wellbeing and liberty following the industrial revolution may be attributable to the fact that human labor was made more productive, being a complement to industrial machinery, and that educated free labor became especially productive in knowledge economies. National power and

wealth increasingly depended on having an educated, free, supportive citizenry. AI could change this 200-year trend if it drastically reduces the value of human labor (by substituting more than it complements labor) and if it reduces governmental authorities' need for the support of their citizens.

Much more work is needed to understand the mechanisms and risks from value erosion. Given that value erosion operates gradually, it may be easier than acute catastrophic risks to observe and coordinate to manage. However, its slow operation and gradual entrenchment of values may also mean that sufficient attention is not mobilized in time.

Early in the evolution of flight, in the late 1920s, some military analysts came to believe that unstoppable bombers dropping poison gas over cities would fundamentally alter warfare (Zaidi, 2021, p. 65). These forecasts were mistaken about the timing and the mechanism of destruction, but they correctly foresaw what would become the strategic logic of the nuclear era, made real by the discovery of the neutron chain reaction. As expressed by the title of a famous collection of essays, the early proponents of controlling nuclear weapons warned that humanity faced a choice: "One World or None." As a matter of fact, confident predictions of nuclear apocalypse were mistaken. However, perhaps they too got the strategic logic right: increasingly powerful technology and great power competition are ultimately not compatible with the long-run flourishing of humanity.

AI governance involves building institutions to guide the development and deployment of AI to achieve our social goals. AI, however, is not a narrow technology, with limited impacts and affordances. The AI revolution will be more like the industrial revolution, transforming economics, politics, and society. To succeed, the field of AI governance must be comparably expansive and ambitious. The body of thought represented by the chapters in this *Handbook* is a great start.

Acknowledgments

I am grateful to many people for input, inspiration, and support with my thinking. For contributions relevant to this work, in particular, I am grateful to: Joslyn Barnhart Alex Belias, Ajeya Cotra, Noemi Dreksler, Lewis Ho, Charlotte Jander, Ramana Kumar, Jade Leung, Tom Lue, David Manheim, Vishal Maini, Silvia Milano, Luke Muehlhauser, Ken Schultz, Rohin Shah, Toby Shevlane, Robert Trager, Adrian Weller, and especially Markus Anderljung, Miles Brundage, Justin Bullock, Ben Garfinkel, and Anton Korinek. Thanks to Leonie Koessler and Alex Lintz for research assistance. Chapter written May 2022.

Notes

1. Defined here simply as machines capable of sophisticated information processing.
2. These numbers are the median forecast from a survey of ML researchers (Zhang et al., 2021).
3. This section draws from Dafoe (2020).
4. For example, see Future of Life Institute (2016).

References

Acemoglu, D. (2024). Harms of AI. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Allison, G.T. (1969, September). Conceptual models and the Cuban missile crisis. *American Political Science Review 63*(3), 689–718. https://doi.org/10.1017/S000305540025853X.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016, July). Concrete problems in AI safety. *arXiv*. https://arxiv.org/abs/1606.06565.

Anderson, B. (2006). *Imagined communities: Reflections on the origin and spread of nationalism*, rev ed. Verso. (Original work published in 1991).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias risk assessments in criminal sentencing. *ProPublica*, May. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arcuri, G. (2022). The CHIPS for America Act: Why it is necessary and what it does. Center for Strategic & International Studies, January. https://www.csis.org/blogs/perspectives-innovation/chips-america-act-why-it-necessary-and-what-it-does.

Armstrong, S., Bostrom, N., & Shulman, C. (2016, May). Racing to the precipice: A model of artificial intelligence development. *AI & Society 31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y.

Askell, A., Brundage, M., & Hadfield, G. (2019, July). The role of cooperation in responsible AI development. *arXiv*. https://doi.org/10.48550/arXiv.1907.04534.

Banks, J. S., (1990). Equilibrium behavior in crisis bargaining games. *American Journal of Political Science 34*(3), 599–614. https://doi.org/10.2307/2111390.

Barnhart, J., (2022). Emerging technology, prestige motivations and the dynamics of International competition. https://www.joslynbarnhart.com/research-1.

Barrett, S. (2007). *Why cooperate? The incentive to supply global public goods*. Oxford University Press.

Bengio, Y., & Chatila, R. (2021, November). Responsible AI working group report. The Global Partnership on Artificial Intelligence. https://gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2021.pdf.

Boix, C. (2024). AI and the economic and informational foundations of democracy. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press, 2014.

Bostrom, N. (2004). The future of human evolution. In C. Tandy (Ed.), *Death and anti-death: Two hundred years after Kant, fifty years after Turing* (pp. 339–371). Ria University Press, 2004.

Bresnahan, T., (2010). General purpose technologies. In B. H. Hall and N. Rosenberg (Eds.), *Handbook of the Economics of Innovation*, *Volume 2* (761–791). Elsevier.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of Econometrics 65*(1), 83–108. https://doi.org/10.1016/0304-4076(94)01598-T.

Brynjolfsson, E. (2022, January). The Turing trap: The promise & peril of human-like artificial intelligence. Stanford Digital Economy Lab. https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/.

Brynjolfsson, E., Collis, A., & Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences 116*(15), 7250–7255. https://doi.org/10.1073/pnas.1815663116.

Caplan, B. (2008). The totalitarian threat. In N. Bostrom, M. M. Cirkovic and M. J. Rees (Eds.), *Global Catastrophic Risks* (pp. 504–519). Oxford University Press.

Carnegie, A., & Carson, A. (2020). *Secrets in global governance: Disclosure dilemmas and the challenge of international cooperation in world politics*. Cambridge University Press.

Christian, B. (2020). *The alignment problem: How can machines learn human values?* Atlantic Books.

Christiano, P. (2019, March). What failure looks like. *AI Alignment Forum*. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Coe, A. J., & Vaynman, J. (2020). Why arms control is so rare. *American Political Science Review 114*(2). https://doi.org/10.1017/S000305541900073X.

Cotra, A. (2020). Draft report on AI timelines. *LessWrong*, September. https://www.lesswrong.com/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines.

Coyle, D., & Weller, A. (2020). "Explaining" machine learning reveals policy challenges. *Science 368*(6498), 1433–1434. https://doi.org/10.1126/science.aba9647.

Dafoe, A. (2018, August). AI governance: A research agenda. Centre for the Governance of AI. https://www.governance.ai/research-paper/agenda.

Dafoe, A. (2020, September). AI governance: Opportunity and theory of impact. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact.

Dafoe, Allan, Hughes, Edward, Bachrach, Yoram, Collins, Tantum, McKee, Kevin R., Leibo, Joel Z., Larson, Kate, & Graepel, Thore. (2020). Open problems in cooperative AI. *arXiv* preprint arXiv:2012.08630.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature 593*(7857), 33–36.

Danzig, R. (2022, January). Machines, bureaucracies, and markets as artificial intelligences. Center for Security and Emerging Technologies. https://cset.georgetown.edu/publication/machines-bureaucracies-and-markets-as-artificial-intelligences/.

Danzig, R. (2018, May). Technology roulette: Managing loss of control as many militaries pursue technological superiority. Center for a new American Security. https://www.cnas.org/publications/reports/technology-roulette.

Ding, J., & Dafoe, A. (2021, June). Engines of power: Electricity, AI, and general-purpose military transformations. *arXiv*. https://arxiv.org/abs/2106.04338.

Drutman, L. (2020). *Breaking the two-party doom loop: The case for multiparty democracy in America*. Oxford University Press.

Foroohar, Rana. (2019.) *Don't Be Evil: How Big Tech Betrayed Its Founding Principles*. Currency/Crown Publishing Group.

Fearon, J. D. (1995). Rationalist explanations for war. *International Organization 49*(3), 379–414. https://doi.org/10.1017/S0020818300033324.

Frey, Carl Benedikt. (2019). *The technology trap*. Princeton University Press.

Future of Life Institute. (2016). Autonomous weapons: An open letter from AI & robotics researchers. https://futureoflife.org/2016/02/09/open-letter-autonomous-weapons-ai-robotics/.

Garfinkel, B. (2020, March). The case for privacy optimism", *The Best That Can Happen* (blog). https://benmgarfinkel.wordpress.com/2020/03/09/privacy-optimism-2/.

Garfinkel, B. (2024). The impact of artificial intelligence: A historical perspective. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research 62* , 729–754. https://doi.org/10.1613/jair.1.11222.

Gruetzemacher, R., & Whittlestone, J. (2022, January). The transformative potential of artificial intelligence. *Futures 135*, 102884. https://doi.org/10.1016/j.futures.2021.102884.

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the Earth*. Oxford University Press.

Hardt, M., Price, E., & Srebro, N. (2016, October). Equality of opportunity in supervised learning. *arXiv*. https://arxiv.org/abs/1610.02413v1.

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021, December). Unsolved problems in ML safety. *arXiv*. https://doi.org/10.48550/arXiv.2109.13916.

Hernández-Orallo, J. (2017). *The measure of all minds: Evaluating natural and artificial intelligence*. Cambridge University Press.

Horowitz, M., Pindyck, S., & Mahoney, C. (2024). AI, the international balance of power, and national security strategy. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Horowitz, M., Kania, E. B., Allen, G. C., & Scharre P. (2018, July). Strategic competition in an era of artificial intelligence. Center for a New American Security. https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence.

Hunt, W. (2020, August). The flight to safety-critical AI: Lessons in AI safety from the aviation industry. UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/2020/08/11/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry/.

Hwang, T. (2020, June). Shaping the terrain of AI competition. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/shaping-the-terrain-of-ai-competition/.

Irving, G., & Askell, A. (2019, February). AI safety needs social scientists. *Distill*. https://distill.pub/2019/safety-needs-social-scientists/.

Jones, C.I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review 110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330.

Karnofsky, H. (2016, May). Potential risks from advanced artificial intelligence: The philanthropic opportunity. Open Philanthropy Project. http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity.

Karnofsky, H. (2021, November). Did life get better during the pre-industrial era? (Ehhhh). *Cold Takes*. https://www.cold-takes.com/did-life-get-better-during-the-pre-industrial-era-ehhhh/.

Kelly, K. (2014, October). The three breakthroughs that have finally unleashed AI on the world. *Wired*. https://www.wired.com/2014/10/future-of-artificial-intelligence/.

Koremenos, B., Lipson, C., & Snidal, D. (2001). The rational design of international institutions. *International Organization 55*(4), 761–799. http://www.jstor.org/stable/3078615.

Korinek, A., & Juelfs, M. (2024). Preparing for the (non-existent?) future of work. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Korinek, A., & Stiglitz, J. (2019). Artificial intelligence and its implications for income distribution and unemployment. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence* (pp. 349–390). University of Chicago Press.

Kristensen, M., Korda, M., & Norris, R. (2022, February). Status of world nuclear forces. Federation of American Scientists. https://fas.org/issues/nuclear-weapons/status-world-nuclear-forces/.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review 165*(3), 633–706. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

LaPointe, C., & Levin, P.L. (2016). Automated war: How to think about intelligent autonomous systems in the military. *Foreign Affairs*, September. https://www.foreignaffairs.com/articles/2016-09-05/automated-war.

Lipsey, R. G., Carlaw K. I., & Bekar, C. T. (2005). *Economic transformations: General purpose technologies and long-term economic growth*. Oxford University Press.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Luttwak, E. N. (1968). *Coup d'état: A practical handbook*. Allen Lane.

Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy 40*(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464.

Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). *Algorithmic game theory*. Cambridge University Press.

Nitzan, S. (1994). Modelling rent-seeking contests. *European Journal of Political Economy 10*(1), 41–60. https://doi.org/10.1016/0176-2680(94)90061-2.

North, D. C. (1991). Institutions. *Journal of Economic Perspectives 5*(1), 97–112. http://www.jstor.org/stable/1942704.

Obama, B. (2020). *A Promised land: Barack Obama*. Viking.

Polachek, S., & Xiang, J. (2010). How opportunity costs decrease the probability of war in an incomplete information game. *International Organization 64*(1), 133–144. http://www.jstor.org/stable/40607983.

Polyakova, A., & Meserole, C. (2019, August). Exporting digital authoritarianism: The Russian and Chinese models. The Brookings Institution. https://www.brookings.edu/research/exporting-digital-authoritarianism/.

Prunkl, C., & Whittlestone, J. (2020). Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February), 138–143. https://doi.org/10.1145/3375627.3375803.

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Scharre, P. (2018, September). A million mistakes a second. *Foreign Policy*. https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

Scharre, P. (2021). Debunking the AI arms race theory. *Texas National Security Review 4*(3), 121–132. http://dx.doi.org/10.26153/tsw/13985.

Scott, J. C. (2008). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Sen, A., Fitoussi, J. P., & Stiglitz, J. (2010). *Mismeasuring our lives: Why GDP doesn't add up*. The New Press.

Somers, J. (2017, April). Torching the modern-day library of Alexandria. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/.

Stockholm International Peace Research Institute. (n.d.). SIPRI military expenditure database. https://www.sipri.org/databases/milex.

Tannenwald, N. (1999). The nuclear taboo: The United States and the normative basis of Nuclear non-use. *International Organization 53*(3), 433–468. https://doi.org/10.1162/002081899550959.

Whittlestone, J., & Clarke, S. (2024). AI challenges for society and ethics. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2019). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4*(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.

Yudkowsky, E. (2008). Artificial Intelligence as a positive and negative factor in global risk. In N. Bostrom and M. M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308–345). Oxford University Press.

Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., & Weller, A. (2017, November). From Parity to preference-based notions of fairness in classification. *arXiv*. https://arxiv.org/abs/1707.00010.

Zaidi, W. H. (2021). *Technological internationalism and world order aviation, atomic energy, and the search for international peace, 1920–1950*. Cambridge University Press.

Zaidi, W., & Dafoe, A. (2021, March). International control of powerful technology: Lessons from the Baruch plan for nuclear weapons. Centre for the Governance of AI. https://www.governance.ai/research-paper/international-control-of-powerful-technology-lessons-from-the-baruch-plan-for-nuclear-weapons.

Zhang, B., Dreksler, N., Anderljung, M., Kahn, L., Giattino, C., Dafoe, A., & Horowitz, M. C. (2021). Forecasting AI progress: Evidence from a survey of machine learning researchers. https://osf.io/v7f6g/. Retrieved November 2021.

Zwetsloot, R., & Dafoe, A. (2019, February). Thinking about risks from AI: Accidents, misuse and structure. *Lawfare* (blog). https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

Zwetsloot, R., Toner, H., & Ding, J. (2018). Beyond the AI arms race: America, China, and the dangers of zero-sum thinking. *Foreign Affairs*, November. https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race.

CHAPTER 2

AI CHALLENGES FOR SOCIETY AND ETHICS

JESS WHITTLESTONE AND SAMUEL CLARKE

Introduction

AI is already being applied in and impacting many important sectors in society, including healthcare, finance, and policing. As investment into AI research continues, we are likely to see substantial progress in AI capabilities and their potential applications, precipitating even greater societal impacts. The use of AI promises real benefits by helping us to better understand the world around us and develop new solutions to important problems, from disease to climate change. However, the power of AI systems also means that they risk causing serious harm if misused or deployed without careful consideration for their immediate and wider impacts.[1]

The role of AI governance is ultimately to take practical steps to mitigate this risk of harm while enabling the benefits of innovation in AI. To do this requires answering challenging empirical questions about the possible risks and benefits of AI, as well as challenging normative questions about what the beneficial use of AI in society looks like.

To properly assess risks and benefits, we need a thorough understanding of how AI is already impacting society, and how those impacts are likely to evolve in future—which is far from straightforward. Assessing even *current* impacts of a technology like AI is challenging because these are likely to be widely and variably distributed across society. Furthermore, it is difficult to determine the extent to which impacts are caused by AI systems, as opposed to other technologies or societal changes. Assessing *potential* impacts of AI in the future—which is necessary if we are to intervene while impacts can still be shaped and harms have not yet occurred—is even more difficult because it requires making predictions about an uncertain future.

The normative question of what beneficial use of AI in society looks like is also complex. A number of different groups and initiatives have attempted to articulate and agree on high-level principles that uses of AI should respect, such as privacy, fairness, and autonomy (Jobin et al., 2019). Though this is a useful first step, many challenges arise when putting these principles into practice. For example, it seems straightforward to say that the

use of AI systems must protect individual privacy, but there is presumably some amount or type of privacy that most people would be willing to give up to develop life-saving medical treatments. Different groups and cultures will inevitably have different views on what trade-offs we should make, and there may be no obvious answer or clear way of adjudicating between views. We must therefore also find politically feasible ways to balance different perspectives and values in practice, and ways of making decisions about AI that will be viewed as legitimate by all.

Despite these challenges, research can and has made progress on understanding the impacts of AI, and on illuminating the challenging normative questions that these impacts raise. The aim of this chapter will be to give the reader an understanding of this progress, and the challenges that remain. We begin by outlining some of the benefits and opportunities AI promises for society, before turning to some of the most concerning sources of harm and risk AI might pose. We then discuss the kinds of ethical and political challenges that arise in trying to balance these benefits and risks, before concluding with some recommendations for AI governance today.

Benefits and Opportunities

The promise of AI ultimately lies in its potential to help us understand the world and solve problems more effectively than humans could do alone. We discuss potential benefits of AI in three related categories: (1) improving the quality and length of people's lives, (2) improving our ability to tackle problems as a society, (3) enabling moral progress and cooperation.

Improving the quality and length of people's lives

AI can help improve the quality and efficiency of public services and products by tailoring them to a given person or context. For example, several companies have begun to use AI to deliver personalized education resources (Hao, 2019), collecting data on students' learning and performance and using this to better understand learning patterns and specific learning needs (Luan & Tsai, 2021). Similarly, the use of AI to personalize healthcare through precision medicine—i.e., tailoring treatment based on specific features of an individual patient—is in early stages but shows real promise (Xu et al., 2019; Johnson et al., 2021), with startups beginning to emerge in this space (Toews, 2020).

AI is also showing promise to drastically improve our understanding of disease and medical treatments. AI systems can now outperform human specialists on a number of specific healthcare-related tasks: for example, Google Health trained a model to predict risk of breast cancer from mammograms, which outperformed human radiologists (McKinney et al., 2020). The use of AI to advance drug discovery, for instance by searching through and testing chemical compounds more quickly and effectively, is receiving increasing attention (Paul et al., 2021): the first clinical trial of an AI-designed drug began in Japan (Burki, 2020) and a number of startups in this space raised substantial funds in 2020 (Hogarth & Benaich, 2020). DeepMind's AI system AlphaFold has led to substantial progress on the "protein folding" problem,[2] with potential to drastically improve our ability to treat disease (Jumper

et al., 2021). Continued progress in AI for healthcare might even contribute to better understanding and slowing processes of aging (Zhavoronkov et al., 2019), resulting in much longer lifespans than we enjoy today.

Improving our ability to tackle problems as a society

AI could help tackle many of the big challenges we face as a society, such as climate change and threats to global health, by helping model the complex systems underpinning these problems, advancing the science behind potential solutions, and improving the effectiveness of policy interventions.

For instance, AI can support early warning systems for threats such as disease outbreaks: machine learning (ML) algorithms were used to characterize and predict the transmission patterns of both Zika (Jiang et al., 2018) and SARS-CoV-2 (Wu et al., 2020; Liu, 2020) outbreaks, supporting more timely planning and policymaking. With better data and more sophisticated systems in future it may be possible to identify and mitigate such outbreaks much earlier (Schwalbe & Wahl, 2020). There is also some early discussion of how AI could also be used to identify early signs of inequality and conflict. Musumba et al. (2021), for instance, use machine learning to predict the occurrence of civil conflict in Sub-Saharan Africa. This could make it much easier to intervene early to prevent conflict.

AI-based modeling of complex systems can improve resource management, which may be particularly important in mitigating the effects of climate change. For instance, AI is beginning to see application in predicting day-ahead electricity demand in the grid, improving efficiency, and in learning how to optimally allocate resources such as fleets of vehicles to address constantly changing demand (Hogarth & Benaich, 2019). Similarly, a better understanding of supply and demand in electricity grids can also help reduce reliance on high-polluting plants, and make it easier to proactively manage an increasing number of variable energy sources (Rolnick et al., 2019). Similar kinds of analysis could help with a range of other problems, including disaster response: for example, machine learning can be used to create maps from aerial imagery and retrieve information from social media to inform relief efforts (Rolnick et al., 2019).

AI also has potential to advance science in critical areas. There are many ways that AI could improve different aspects of the scientific process: by helping us to understand and visualize patterns in data of enormous volume and dimensionality (Mjolsness & DeCoste, 2001; Ourmazd, 2020); or by conducting more "routine" aspects of scientific research such as literature search and summarization, hypothesis generation, and experimental design and analysis (Gil et al., 2014). DeepMind's work on protein folding mentioned earlier is a good example of AI already advancing science in an important area. In the future, we could see AI accelerating progress in areas like materials science, by automating the time-consuming processes in the discovery of new materials, which could help develop better materials for storing or harnessing energy, for example (Rolnick et al., 2019).

As well as improving our understanding of problems and advancing the science needed to solve them, AI can help identify the most effective solutions that currently exist. There is evidence that ML tools can be used to improve policymaking by clarifying uncertainties in data, and improving existing tools for designing and assessing interventions (Rolnick et al., 2019). For instance, Andini et al. (2018) show that a simple ML algorithm could have been

used to increase the effectiveness of a tax rebate program. It may even be possible to use AI to design more competent institutions which would help tackle many problems. One idea here is that (human) participants could determine desiderata that some institution should achieve, and leave the design of the institution to an AI system (Dafoe et al., 2020). This could allow novel approaches to old problems that humans cannot spot.

Enabling moral progress and cooperation

Most would agree that the world we live in today is a better place for most people than the world of centuries ago. This is partly due to economic and technological progress improving standards of living across the globe. But moral progress also plays an important role. Fewer people and animals experience suffering today, for example, because most people view an increasing proportion of sentient beings as worthy of care and moral concern. It has been suggested that AI could help accelerate moral progress (Boddington, 2021), for example by playing a "Socratic" role in helping us to reach better (moral) decisions ourselves (inspired by the role of deliberative exchange in Socratic philosophy as an aid to develop better moral judgements) (Lara & Deckers, 2020). Specifically, such systems could help with providing empirical support for different positions, improving conceptual clarity, understanding argumentative logic, and raising awareness of personal limitations.

AI might similarly help improve cooperation between groups, which arguably underlies humans' success in the world so far. Dafoe et al. (2020) outline a number of ways AI might support human cooperation: AI tools could help groups jointly learn about the world in ways that make it easier to find cooperative strategies, and more advanced machine translation could enable us to overcome practical barriers to increased international cooperation, including increased trade and possibly leading to a more borderless world. AI could also play an important role in building mechanisms to incentivize truthful information sharing, and in exploring the space of distributed institutions that promote desirable cooperative behaviors.

Harms and Risks

Despite these many real and potential benefits, we are already beginning to see harms arise from the use of AI systems, which could become much more severe with more widespread application of increasingly capable systems.

In this section, we'll discuss five different forms of harm AI might pose for individuals and society, in each case outlining current trends and impacts of AI pointing in this direction, and what we might be especially concerned about as AI systems increase in their capabilities and ubiquity across society.

Increasing the likelihood or severity of conflict

AI could impact the severity of conflict by enabling the development of new and more lethal weapons. Of particular concern are lethal autonomous weapons (LAWs): systems that

can select and engage targets without further intervention by a human operator, which may recently have been used in combat for the first time (UN Security Council, 2021).[3] There is a strong case that "armed fully autonomous drone swarms," one type of lethal autonomous weapon, qualify as a weapon of mass destruction (WMD) (Kallenborn, 2020). This means they would pose all the threats that other WMDs do: geopolitical destabilization and use in acts of terror or catastrophic conflict between major powers. They would also be safer to transport and harder to detect than most other WMDs (Aguire, 2020). Beyond LAWs, AI applied to scientific research or engineering could enable the development of other extremely powerful weapons. For example, it could be used to calculate the most dangerous genome sequences in order to create especially virulent biological viruses (O'Brien & Nelson, 2020; Turchin & Denkenberger, 2020).

Furthermore, we are seeing more integration of AI into defense and conflict domains, which could increase the likelihood of unintentional or rapid escalation in conflict: if more military decisions are automated, this makes it harder to intervene to prevent escalation (Johnson, 2020; Deeks et al., 2018). This is analogous to how algorithmic decision-making in financial systems led to the 2010 "flash crash:" automated trading algorithms, operating without sufficient oversight, caused a trillion-dollar stock market crash over a period of approximately 36 minutes. The consequences could be even worse in a conflict scenario than in finance, because there is no overarching authority to enforce failsafe mechanisms (Johnson, 2020).

AI could also alter incentives in a way that makes conflict more likely to occur or to escalate (Zwetsloot & Dafoe, 2019). For example, AI could undermine second strike capabilities which are central to nuclear strategic stability, by improving data collection and processing capabilities which would make it easier to discover and destroy previously secure nuclear launch facilities (Geist & Lohn, 2018; Lieber & Press, 2017).

Making society more vulnerable to attack or accident

As AI systems become more integral to the running of society this may create new vulnerabilities which can be exploited by bad actors. For instance, researchers managed to fool an ML model trained to recognize traffic signs into classifying a "stop" sign as a "yield" sign, simply by adding a small, imperceptible perturbation to the image (Papernot et al., 2017). An autonomous vehicle using this model could therefore be targeted by bad actors using stickers or paint to alter traffic signs. As AI systems become more widely deployed, these kinds of attacks could have more catastrophic consequences. For example, as AI is more widely integrated into diagnostic tools in hospitals or into our transport systems, adversarial attacks could put many lives at risk (Finlayson et al., 2019; Brundage et al., 2018).

Similarly, more widespread deployment of increasingly capable AI systems could also increase the severity of accidents. In particular, although the integration of AI into critical infrastructure has potential to bring efficiency benefits, it would also introduce the possibility of accidents on a far more consequential scale than is possible today. For example, as driverless cars become more ubiquitous, computer vision systems failing in extreme weather or road conditions could cause many cars to crash simultaneously. The direct casualties and second-order effects on road networks and supply chains could be severe. If and when AI systems become sufficiently capable to run large parts of society, these kinds of failures

could possibly result in the malfunction of several critical systems at once, which at the extreme could put our very civilization at risk of collapse.

One might think that these accidents could be avoided by making sure that a human either approves or makes the final decision. However, progress in AI capabilities such as deep reinforcement learning (DRL) could lead us to develop more autonomous systems, and there will likely be commercial pressure to deploy them. For such systems, especially when their decisions are too fast-moving or incomprehensible to humans, it is not clear how human oversight would work (Whittlestone et al., 2021).

These risks may be exacerbated by competitive dynamics in AI development. AI development is often framed in terms of a "race" for strategic advantage and technological superiority between nations (Cave & ÓhÉigeartaigh, 2018). This framing is prominent in news sources, the tech sector, and reports from governmental departments, such as the U.S. Senate and Department of Defense (Imbrie et al., 2020). The more AI development is underpinned by these competitive dynamics, there may be a greater incentive for actors developing in AI to underinvest in the safety and security of their systems to stay ahead.

Increasing power concentration

Several related trends suggest AI may change the distribution of power across society, perhaps drastically.

Absent major institutional reform, it seems plausible that the harms and benefits of AI will be very unequally distributed across society. AI systems are already having discriminatory impacts on marginalized groups: for example, facial recognition software has been shown to perform many times worse for darker faces (Raji & Buolamwini, 2019), and an AI system developed by Amazon to rank job candidates downgraded applications whose CVs included evidence they were female (West et al., 2019). Marginalized groups are less technologically literate on average, so are also more likely to be impacted by harms of AI, such as the scaling up of misinformation and manipulative advertising (Lutz, 2019). These groups are also less likely to be in a financial position to benefit from advances in AI (e.g., personalized healthcare) (West et al., 2019).

At the same time, AI development is making already wealthy and powerful actors more so. The companies who already have the greatest market share have access to the most data, computing power, and research talent, enabling them to build the most effective products and services—increasing their market share further and making it easier for them to continue amassing data, compute, and talent (Dafoe, 2018; Kalluri, 2020; Lee, 2018). This creates a positive feedback loop cementing the powerful position these technology companies are already in. Similarly, wealthier countries able to invest more in AI development are likely to reap economic benefits more quickly than developing economies, potentially widening the gap between them. Especially if AI development leads to more rapid economic growth than previous technologies (Aghion et al., 2019), this might cause more extreme concentration of power than we have ever seen before.

In addition, AI-based automation has the potential to drastically increase income inequality. Progress in AI systems will inevitably make it possible to automate an increasing range of tasks. Progress in reinforcement learning specifically could improve the dexterity and flexibility of robotic systems (Ibarz et al., 2021), leading to increased automation of

manual labor jobs with lower wages. The automation of these jobs will force those people to retrain; even in the best case, they will face temporary disruptions to income (Lee, 2018). However, it is not just low-wage or manual labor jobs that are at risk. Advances in language modeling could spur rapid automation of a wide range of knowledge work, including aspects of journalism, creative writing, and programming (Tamkin et al., 2021). Many of these knowledge workers will flood the highly social and dextrous job market (which is hard to automate, but already has low wages), further increasing income inequality (Lee, 2018). There is also reason to think that changes in the availability of jobs due to AI may happen more quickly than previous waves of automation, due to the fact that algorithms are infinitely replicable and instantly distributable (unlike, for example, steam engines and even computers), and the emergence of highly effect venture capital funding driving innovation (Lee, 2018). This gives us less time to prepare, for example by retraining those whose jobs are most likely to be lost, and it makes it more likely that the impacts on inequality will be more extreme than anything seen previously.

Developments in AI are also likely to give companies and governments more control over individuals' lives than ever before. The fact that current AI systems require large amounts of data to learn from creates incentives for companies to collect increasing amounts of personal data from users (though only certain applications, such as medicine and advertising, require highly personal data). Citizens are increasingly unable to consent to—or even be aware of—how their data is being used, while the collection of this data may increasingly be used by powerful actors to surveil, influence, and even manipulate and control populations. For example, the company ClearView AI scraped billions of images from Facebook, YouTube, Venmo, and millions of other websites, using them to develop a "search engine for faces," which they then licensed, without public scrutiny, to over 600 law enforcement agencies (Hill, 2020). We are already seeing harmful uses of facial recognition, such as in their use to surveil Uighur and other minority populations in China (Hogarth & Benaich, 2019).[4] The simultaneous trends of apparently eroding privacy norms, and increased use of AI to monitor and influence populations, are seriously concerning.

Relatedly, AI has the potential to scale up the production of convincing yet false or misleading information online (e.g., via image, audio and text synthesis models like BigGAN and GPT-3), and to target that content at individuals and communities most likely to be receptive to it (e.g., via automated A/B testing) (Seger at al., 2020). Whilst the negative impact of such techniques has so far been fairly contained, more advanced versions would make it easier for groups to seek and retain influence, for instance by influencing elections or enabling highly effective propaganda. For example, further advances in language modeling could be applied to design tools that "coach" their users to persuade other people of certain claims (Kokotajlo, 2020). Whilst these tools could be used for social good—e.g., *The New York Times*' chatbot that helps users to persuade people to get vaccinated against COVID-19 (Gagneur & Tamerius, 2021)—they could equally be used by self-interested groups to gain or retain influence.

Undermining society's ability to solve problems

The use of AI in the production and dissemination of information online may also have broader negative impacts. In particular, it has been suggested that the use of AI to improve

content recommendation engines by social media companies is contributing to worsened polarization online (Ribeiro et al., 2019; Faddoul et al., 2020).[5]

Looking to the future, the use of AI in production or targeting of information could have substantial impacts on our information ecosystem. If advanced persuasion tools are used by many different groups to advance many different ideas, we could see the world splintering into isolated "epistemic communities," with little room for dialogue or transfer between them. A similar scenario could emerge via the increasing personalization of people's online experiences: we may see a continuation of the trend towards "filter bubbles" and "echo chambers," driven by content selection algorithms, that some argue is already happening (Barberá et al., 2015; Flaxman et al., 2016; Nguyen et al., 2014). In addition, increased awareness of these trends in information production and distribution could make it harder for anyone to evaluate the trustworthiness of any information source, reducing overall trust in information.

In all these scenarios, it would be much harder for humanity to make good decisions on important issues, particularly due to decreasing trust in credible multipartisan sources, which could hamper attempts at cooperation and collective action. The vaccine and mask hesitancy which exacerbated the negative impacts of COVID-19, for example, were likely the result of insufficient trust in public health advice (Seger, 2021). We could imagine an even more virulent pandemic, where actors exploit the opportunity to spread misinformation and disinformation to further their own ends. This could lead to dangerous practices, a significantly increased burden on health services, and much more catastrophic outcomes (Seger et al., 2020).

Losing control of the future to AI systems

If AI systems continue to become more capable and begin running significant parts of the economy, we might also worry about humans losing control of important decisions. Currently, humans' attempts to shape the world are the only goal-directed process influencing the future. However, more automated decision-making would change this, and could result in some (or all) human control over the future being lost (Christiano, 2019; Critch, 2021; Ngo, 2020; Russell, 2019).

This concern relies on two assumptions. First, that AI systems will become capable enough that it will be not only possible but desirable to automate a majority of tasks making up the economy, from managing critical infrastructure to running corporations. Second, that despite our best efforts, we may not understand these systems well enough to be sure they are fully aligned with what their operators want.

How plausible are these assumptions? Considering the first, there is increasing reason to believe we might build AI systems as capable as humans across a broad range of economically useful tasks this century. Enormous amounts of resources are going into AI progress and developing human-level AI is the stated goal of two very well-resourced organizations (DeepMind and OpenAI), as well as a decent proportion of AI researchers. In recent years, we have seen advances in AI defy expectations, especially in terms of their ability to solve tasks they weren't explicitly trained for, and the improvements in performance that can be derived from simply increasing the size of models, the datasets they are trained on, and the computational resources used for training them (Branwen, 2021).[6] For example, GPT-3 (the

latest language model from OpenAI at the time of writing), shows remarkable performance on a range of tasks it was not explicitly trained on, such as generating working code from natural language descriptions, functioning as a chatbot in limited contexts, and being used as a creative prompt (Tamkin et al., 2021). These capabilities are quickly spurring a range of commercial applications, including GitHub Copilot, a tool that helps programmers work faster by suggesting lines of code or entire functions (Chen et al., 2021). This progress was achieved simply by scaling up previous language models to larger sizes and training them with more data and computational resources. There is good evidence that this trend will continue to result in more powerful systems without needing "fundamental" breakthroughs in machine learning (Kaplan et al., 2020).

The second assumption, that advanced AI systems might not be fully aligned with or understandable to humans, is perhaps on even stronger ground. We currently train AI systems by "trial and error," in the sense that we search for a model that does well on some objective, without necessarily knowing how a given model produces the behavior it does. This leaves us with limited assurance about how the system might behave in new contexts or environments. A particular concern is that AI systems might help us to optimize for what we can *measure* in society, but not what we actually value (Christiano, 2019). For example, we might deploy AI systems in law enforcement to help increase security and safety in communities, but later find that these systems are in fact increasing a *reported* sense of safety by driving down complaints and hiding information about failures. If we don't notice these kinds of failures until AI systems are integral to the running of society, it may be very costly or even impossible to correct them. As mentioned earlier, competitive pressures to use AI for economic gain may make this more likely, driving actors to deploy AI systems without sufficient assurances that they are optimizing for what we want.

This could happen gradually or suddenly, depending on the pace and shape of AI progress. The most high-profile versions of these concerns have focused on the possibility of a single misaligned AI system rapidly increasing in intelligence (Bostrom, 2014), but a much more gradual "takeover" of society by AI systems may be more plausible, where humans don't quite realize they are losing control until society is almost entirely dependent on AI systems and it is difficult or impossible for humans to regain control over decision-making.

Ethical and Political Challenges

It is fairly uncontroversial to suggest that improving the length and quality of people's lives is something we should strive for, and that catastrophic accidents which claim thousands of lives should be avoided.

However, enabling the benefits of AI while mitigating the harms is not necessarily so straightforward. Sometimes what is needed to enable some area of benefit may also be the exact thing that carries risk.

For example, using AI to automate increasing amounts of the economy has potential to improve the quality of services and rapidly boost economic growth, which could result in drastic improvements to quality of life across the globe. However, the economic gains

of this kind of progress, as well as the harms of job displacement, may be drastically unequal, leading to a concentration of power and rise in inequality across society never seen before. There are empirical questions here, about what processes are most likely to exacerbate inequality, that research could make progress on. There are also practical interventions that could be implemented to increase the likelihood that the economic gains of AI can be redistributed. However, there are still fundamental value judgements that must be made when envisioning what we want from the future of AI: how should we balance the potential for societal progress, and the possibility of huge gains in average quality of life, against the risk of radically increased inequality? If applying AI to science has potential to increase human health and lifespans, but also risks the creation of dangerous new technologies if not approached with care and wisdom, how much risk should we be willing to take? If outsourcing decisions to AI systems has potential to help us solve previously intractable societal problems, but at the cost of reduced human autonomy and understanding of the world, what should we choose?

Because these questions are normatively complex, there will be plenty of room for reasonable disagreement. Those who prioritize aggregate wellbeing will want to make different choices today to those who prioritize equality. Younger people may be happier to sacrifice privacy than older generations; those from countries which already have a strong welfare state will likely be more concerned about threats to equality; and values such as human autonomy may be perceived very differently in different cultures.

How do we deal with these disagreements? In part, this is the domain of AI ethics research, which can help to illuminate important considerations and clearly outline arguments for different perspectives. However, we should not necessarily expect ethics research to provide all the answers, especially on the time frame in which we need to make decisions about how AI is developed and used. We can also provide opportunities for debate and resolution but, in most cases, it will be impossible to resolve disagreements entirely and use AI in ways everyone agrees with.[7] We must therefore find ways to make choices about AI despite the existence of complex normative issues and disagreement on them.

Some political scientists and philosophers have suggested that where agreement on final decisions is impossible, we should instead focus our attention on ensuring the *process* by which a decision is made is legitimate (Patty & Penn, 2014). This focus on decision-making procedures as opposed to outcomes has also arisen in debates around public health ethics. Daniels and Sabin (2008) suggest that in order to be seen as legitimate, decision-making processes must be, among other things, open to public scrutiny, revision, and appeal.[8]

We do not currently have legitimate procedures for making decisions about how we develop and use AI in society. Many important decisions are being made in technology companies whose decisions are not open to public or even government scrutiny, meaning they have little accountability for the impacts of their decisions on society. For instance, despite being among the world's most influential algorithms, Facebook's and YouTube's content selection algorithms are mostly opaque to those most impacted by them. The values and perspectives of individuals making important decisions have disproportionate influence over how "beneficial AI" is conceived of, while the perspectives of minority groups and less powerful nations have little influence.

Implications for Governance

What should we be doing to try to ensure that AI is developed and used in beneficial ways, today and in the future? We suggest that AI governance today should have three broad aims.

Ultimately, AI governance should be focused on *identifying and implementing mechanisms which enable benefits and mitigate harms of AI*. However, as we've discussed throughout this chapter, in some cases doing this may not be straightforward, for two reasons. First, there are many actual and potential impacts of AI which we do not yet understand well enough to identify likely harms and benefits. Second, even where impacts are well understood, tensions may arise, raising challenging ethical questions on which people with different values may disagree. AI governance therefore also needs to develop methods and processes to address these barriers: *to improve our ability to assess and anticipate the impacts of AI, and to make decisions even in the face of normative uncertainty and disagreement*. We conclude this chapter by making some concrete recommendations for AI governance work in each of these three categories.

Enabling benefits and mitigating harms

In some cases, *we might need to consider outright bans on specific applications of AI*, if the application is likely to cause a level or type of harm deemed unacceptable. For example, there has been substantial momentum behind campaigns to ban lethal autonomous weapons (LAWs), and the European Commission's proposal for the first AI regulation includes a prohibition on the use of AI systems which engage in certain forms of manipulation, exploitation, indiscriminate surveillance, and social scoring (European Commission, 2021). Another area where prohibitions may be appropriate is in the integration of AI systems into nuclear command and control, which could increase the risk of accidental launch with catastrophic consequences, without proportional benefits (Ord et al., 2021).

However, effective bans on capabilities or applications can be challenging to enforce in practice. It can be difficult to achieve the widespread international agreement needed. For example, the U.S. government has cited the fact that China is unlikely to prohibit LAWs as justification for not making the ban themselves (NSCAI, 2021). In other cases, it may be difficult to delineate harmful applications clearly enough. In the case of the EU regulation, it is likely to be very difficult to clearly determine whether an AI system should be deemed as "manipulative" or "exploitative" in the ways stated, for example.

Where outright bans are infeasible, *it may be possible to limit access to powerful capabilities to reduce risk of misuse*. For example, companies might choose not to publish the full code behind specific capabilities to prevent malicious actors from being able to reproduce them, or limit access to commercial products with potential for misuse (Radford et al., 2019). However, this introduces a tension between the need for caution and the benefits of open sharing in promoting beneficial innovation (Whittlestone & Ovadya, 2020), which has prompted substantial debate and analysis around the role of publication norms in AI (Gupta et al., 2020). Governments might also consider monitoring and regulating access

to large amounts of computing power, which would allow them oversight and control over which actors have access to more powerful AI systems (Brundage et al., 2018).

To go beyond preventing harms and realize the full benefits of AI, it will be crucial to *invest in both socially beneficial applications, and in AI safety and responsible AI research*. Many of the potential benefits we discussed early in this chapter seem relatively underexplored: the potential uses of AI to enhance cooperation between groups, to combat climate change, or to improve moral reasoning, for example, could receive a great deal more attention. Part of the barrier to working on these topics is that they may not be well-incentivized by either academia (which often rewards theoretical progress over applications) or industry (where economic incentives are not always aligned with broad societal benefit). Similarly, AI safety and responsible AI research will be crucial for ensuring even the most beneficial applications of AI do not come with unintended harms. One concrete idea would be for governments to create a fund of computational resources which is available free of charge for projects in these areas (Brundage et al., 2020).

Improving our ability to assess and anticipate impacts

In many cases we may first need to better understand the potential impacts of AI systems before determining what kinds of governance are needed. Better and more standardized processes for impact assessment would be valuable on multiple levels.

First, we need to *establish clearer standards and methods for assuring AI systems* (also sometimes called test, evaluation, validation, and verification—TEVV—methods) before they go to market, particularly in safety-critical contexts. There are currently no proven effective methods for assuring the behavior of most AI systems, so much more work is needed (Flournoy et al., 2020). It is likely that rather than a single approach to assuring AI systems, an ecosystem of approaches will be needed, depending on the type of AI system, and the decision to be made (Ahamat et al., 2021). Better assurance processes would make it easier to decide where the use of AI systems should be restricted, by requiring uses of AI to pass certain established standards. It would also make it possible to identify and mitigate potential harms from unintended behavior in advance, and to incentivize technical progress to make systems more robust and predictable.

Continual monitoring and stress-testing of systems will also be important, given it may not be possible to anticipate all possible failure modes or sources of attack in advance of deployment. Here it may be useful to build on approaches to "read-teaming" in other fields including information and cyber security (Brundage et al., 2018).

We also need *broader ways to assess and anticipate the structural impacts of AI systems*. Assurance and stress-testing can help to identify where unintended behaviors or attacks on AI systems might cause harm, but cannot identify where a system behaving as intended might nonetheless cause broader structural harms (for example, polarizing online discourse or changing incentives to make conflict more likely) (Zwetsloot & Dafoe, 2019). This will likely require looking beyond existing impact assessment frameworks and drawing on broader perspectives and methodologies, including: social science and history, fields which study how large societal impacts may come about without anyone intending them (Zwetsloot & Dafoe, 2019); foresight processes for considering the future evolution

of impacts (Government Office for Science, 2017); and participatory processes to enable a wider range of people to communicate harms and concerns (Smith et al., 2019).

More systematic monitoring of AI progress would improve our ability to anticipate and prepare for new challenges before they arise (Whittlestone & Clark, 2021). As technologies advance and more AI systems are introduced into the market, they will raise increasingly high-stakes policy challenges, making it increasingly important that governments have the capacity to react quickly. AI as a sector is naturally producing a wide range of data, metrics and measures that could be integrated into an "early warning system" for new capabilities and applications which may have substantial impacts on society. Monitoring progress on widely studied benchmarks and assessment regimes in AI could enable AI governance communities to identify areas where new or more advanced applications of AI may be forthcoming. Monitoring *inputs* into AI progress, such as computational costs, data, and funding, may also help to give a fuller picture of where societally-relevant progress is most likely to emerge (Martínez-Plumed et al., 2018). For example, early warning signs of recent progress in language models could have been identified via a combination of monitoring progress on key benchmarks in language modeling and monitoring the large jumps in computational resources being used to train these models.

Making decisions under uncertainty and disagreement

Even with better methods for assessing and anticipating the impacts of AI systems, challenges will remain. There will be uncertainties about the future impacts of AI that cannot be reduced, and conflicting perspectives on how we *should* be using AI for global benefit that cannot be easily resolved. AI governance will therefore need to grapple with what *processes* for making decisions about AI should look like, given this uncertainty and disagreement.

Greater use of participatory processes in decision-making around AI governance could help with ensuring the legitimacy and public acceptability of decisions and may also improve the quality of the decisions themselves. There is evidence that participatory approaches used in the domain of climate policy lead to both increased engagement and understanding of decisions, and to better decisions (Hügel & Davies, 2020). Various projects have begun to engage a wider variety of perspectives in thinking through governance and societal issues related to AI (Ipsos MORI, 2017; Balaram et al., 2018), but much more could be done, especially in terms of integrating these processes into policymaking. We would also like to see participatory studies focused on concerns and hopes about the *future* of AI rather than just current AI systems because these are more likely to be timely and relevant enough to influence decision-making. Public engagement is of course only one kind of input into decision-making processes, and it must be combined with relevant expert analysis. However, participatory processes can be especially useful for understanding the wider impacts of policies which might be neglected by decision-makers, and for highlighting additional considerations or priorities, and policymaking around AI would benefit from giving them greater attention.

More generally, we need to think about how *processes for making important decisions about AI can be sufficiently open to scrutiny and challenge*. This is particularly difficult given

that some of the most important decisions about the future of AI are being made within technology companies, which are not subject to the same forms of accountability or transparency requirements as governments. Some greater scrutiny may be achieved through regulation requiring greater transparency from companies. It may also be possible to improve transparency and accountability through shifts in norms—if there is enough public pressure, companies may have an incentive to be more transparent—or by improving the capacity of government to monitor company behavior, such as by increasing technical expertise in government and establishing stronger measurement and monitoring infrastructure.

Conclusion

In this chapter, we have outlined some of the possible ways AI could impact society into the future, both beneficial and harmful. Our aim has not been to predict the future, but to demonstrate that the possible impacts are wide-ranging, and that there are things we can do today to shape them. As well as intervening to enable specific benefits and mitigate harms, AI governance must develop more robust methods to assess and anticipate the impacts of AI, and better processes for making decisions about AI under uncertainty and disagreement.

Notes

1. When we talk about AI systems in this chapter, we mean software systems which use machine learning (ML) techniques. ML involves learning from data to build mathematical models which can help us with a variety of real-world tasks, including predicting the likelihood a loan will be repaid based on someone's financial history, translating text between languages, or deciding what moves to take to win at a board game.
2. This is the problem of predicting the 3D structure of a protein from its 2D genetic sequence.
3. According to the UN Security Council (2021) report, "Logistics convoys and retreating HAF [in Libya] were subsequently hunted down and remotely engaged by the unmanned combat aerial vehicles or the lethal autonomous weapons systems such as the STM Kargu-2 . . . programmed to attack targets without requiring data connectivity between the operator and the munition: in effect, a true 'fire, forget and find' capability."
4. CloudWalk Technology, a key supplier to the Chinese government, markets its "fire eye" facial recognition service to pick out "Uighurs, Tibetans and other sensitive groups."
5. Note that this suggestion has been disputed (e.g., Ledwich & Zaitsev, 2019; Boxell et al., 2017). The underlying methodological problem is that social media companies have sole access to the data required to perform a thorough analysis, and they lack incentive to publicize this data or perform the analysis themselves.
6. A similar point is made by Sutton (2019): using general methods like search and learning (rather than specific methods than involve building human knowledge into AI systems) and applying a lot of computation to them, has and will continue to yield the biggest breakthroughs in AI.
7. A number of formal results from social choice theory demonstrate that when there are numerous different preferences and criteria relevant to a decision, only under strong

assumptions can an unambiguously "best" option be found—i.e., in many real-life cases, no such resolution will be possible (Patty & Penn, 2014).

8. Of course, there will also be room for reasonable disagreement about decision-making procedures, but we think there is likely to be less disagreement on this level, than on the level of object level decisions.

Bibliography

Aghion, P., Jones, B. F., & Jones, C. I. (2019). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 237–282). The University of Chicago Press. http://www.nber.org/chapters/c14015.

Aguirre, A. (2020). Why those who care about catastrophic and existential risk should care about autonomous weapons. *EA Forum*. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2.

Ahamat, G., Chang, M., & Thomas, C. (2021). Types of assurance in AI and the role of standards. *Centre for Data Ethics and Innovation Blog*. https://cdei.blog.gov.uk/2021/04/17/134/.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization 156* , 86–102. https://doi.org/10.1016/j.jebo.2018.09.010.

Balaram, B., Greenham, T., & Leonard, J. (2018). *Artificial intelligence: Real public engagement*. https://www.thersa.org/reports/artificial-intelligence-real-public-engagement.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science 26* (10), 1531–1542. https://doi.org/10.1177/0956797615594620.

Boddington, P. (2021). AI and moral thinking: How can we live well with machines to enhance our moral agency? *AI and Ethics 1* (2), 109–111. https://doi.org/10.1007/s43681-020-00017-0.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (First edition). Oxford University Press.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences 114* (40), 10612–10617.

Branwen, G. (2021). *The scaling hypothesis*. https://www.gwern.net/Scaling-hypothesis.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., ÓhÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv:1802.07228 [Cs]*. http://arxiv.org/abs/1802.07228.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *ArXiv:2004.07213 [Cs]*. http://arxiv.org/abs/2004.07213.

Burki, T. (2020). A new paradigm for drug development. *The Lancet Digital Health 2* (5), e226–e227. https://doi.org/10.1016/S2589-7500(20)30088-1.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI race for strategic advantage: Rhetoric and risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. https://doi.org/10.1145/3278721.3278780.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., . . . Zaremba, W. (2021). Evaluating large language models trained on code. *ArXiv:2107.03374 [Cs]*. http://arxiv.org/abs/2107.03374.

Christiano, P. (2019). What failure looks like. *AI Alignment Forum*. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Critch, A. (2021). What multipolar failure looks like, and robust agent-agnostic processes (RAAPs). *AI Alignment Forum*. https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic.

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI. http://www.fhi.ox.ac.uk/govaiagenda.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open problems in cooperative AI. *ArXiv:2012.08630 [Cs]*. http://arxiv.org/abs/2012.08630.

Daniels, N., & Sabin, J. E. (2008). Accountability for reasonableness: An update. *BMJ (Clinical Research Ed.) 337* , a1850. https://doi.org/10.1136/bmj.a1850.

Deeks, A., Lubell, N., & Murray, D. (2018). *Machine learning, artificial intelligence, and the use of force by states* (SSRN Scholarly Paper ID 3285879). Social Science Research Network. https://papers.ssrn.com/abstract=3285879.

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy videos. *ArXiv:2003.03318 [Cs]*. http://arxiv.org/abs/2003.03318.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science 363* (6433), 1287–1289. https://doi.org/10.1126/science.aaw4399.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly 80* (S1), 298–320. https://doi.org/10.1093/poq/nfw006.

Flournoy, M. A., Haines, A., & Chefitz, G. (2020). *Building Trust through Testing*. Building Trust through Testing.

Gagneur, A., & Tamerius, K. (2021, May 20). Opinion | Your friend doesn't want the vaccine. What do you say? *The New York Times*. https://www.nytimes.com/interactive/2021/05/20/opinion/covid-19-vaccine-chatbot.html.

Geist, E., & Lohn, A. J. (2018). *How might artificial intelligence affect the risk of nuclear war?* https://www.rand.org/pubs/perspectives/PE296.html.

Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science 346* (6206), 171–172. https://doi.org/10.1126/science.1259439.

Government Office for Science. (2017). *Government Office for Science annual report: 2016 to 2017*. https://www.gov.uk/government/publications/government-office-for-science-annual-report-2016-to-2017.

Gupta, A., Lanteigne, C., & Heath, V. (2020). Report prepared by the Montreal AI ethics institute (Maiei) on publication norms for responsible ai. *ArXiv:2009.07262 [Cs]*. http://arxiv.org/abs/2009.07262.

Hao, K. (2019). China has started a grand experiment in AI education. It could reshape how the world learns. *MIT Technology Review*. https://www.technologyreview.com/2019/08/02/131198/china-squirrel-has-started-a-grand-experiment-in-ai-education-it-could-reshape-how-the/.

Hill, K. (2020, January 18). The secretive company that might end privacy as we know it. *The New York Times*. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Hogarth, I., & Benaich, N. (2019). *State of AI report 2019*. https://www.stateof.ai/2019.

Hogarth, I., & Benaich, N. (2020). *State of AI report 2020*. https://www.stateof.ai/.

Hügel, S., & Davies, A. R. (2020). Public participation, engagement, and climate change adaptation: A review of the research literature. *WIREs Climate Change 11* (4). https://doi.org/10.1002/wcc.645.

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning; lessons we've learned. *ArXiv:2102.02915 [Cs]*. https://doi.org/10.1177/0278364920987859.

Imbrie, A., Dunham, J., Gelles, R., & Aiken, C. (2020). *Mainframes: A provisional analysis of rhetorical frames in AI*. Center for Security and Emerging Technology. https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/.

Ipsos MORI. (2017). *Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society*. Royal Society. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.

Jiang, D., Hao, M., Ding, F., Fu, J., & Li, M. (2018). Mapping the transmission risk of Zika virus using machine learning models. *Acta Tropica 185* , 391–399. https://doi.org/10.1016/j.actatropica.2018.06.021.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1* (9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, J. (2020). Artificial intelligence in nuclear warfare: A perfect storm of instability? *The Washington Quarterly 43* (2), 197–211. https://doi.org/10.1080/0163660X.2020.1770968.

Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and Translational Science 14* (1), 86–93. https://doi.org/https://doi.org/10.1111/cts.12884.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kallenborn, Z. (2020). *Are drone swarms weapons of mass destruction?* (Future Warfare). U.S. Air Force Center for Strategic Deterrence Studies. https://media.defense.gov/2020/Jun/29/2002331131/-1/-1/0/60DRONESWARMS-MONOGRAPH.PDF.

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature 583* (7815), 169–169. https://doi.org/10.1038/d41586-020-02003-2.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *ArXiv:2001.08361 [Cs, Stat]*. http://arxiv.org/abs/2001.08361.

Kokotajlo, D. (2020). *Persuasion tools: AI takeover without AGI or agency?* AI Alignment Forum. https://www.alignmentforum.org/posts/qKvn7rxP2mzJbKfcA/persuasion-tools-ai-takeover-without-agi-or-agency.

Lara, F., & Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics 13* (3), 275–287. https://doi.org/10.1007/s12152-019-09401-y.

Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *ArXiv:1912.11211 [Cs]*. http://arxiv.org/abs/1912.11211.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Lieber, K. A., & Press, D. G. (2017). The new era of counterforce: Technological change and the future of nuclear deterrence. *International Security 41* (4), 9–49. https://doi.org/10.1162/ISEC_a_00273.

Liu, J. (2020, April 2). *Deployment of health IT in China's fight against the COVID-19 pandemic*. Imaging Technology News. https://www.itnonline.com/article/deployment-health-it-china%E2%80%99s-fight-against-covid-19-pandemic.

Luan, H., & Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society 24* (1), 250–266.

Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies 1* (2), 141–148. https://doi.org/https://doi.org/10.1002/hbe2.140.

Martínez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., ÓhÉigeartaigh, S., & Hernández-Orallo, J. (2018). Accounting for the neglected dimensions of AI progress. *ArXiv:1806.00610 [Cs]*. http://arxiv.org/abs/1806.00610.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature 577* (7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Mjolsness, E., & DeCoste, D. (2001). Machine learning for science: State of the art and future prospects. *Science 293* (5537), 2051–2055. https://doi.org/10.1126/science.293.5537.2051.

Musumba, M., Fatema, N., & Kibriya, S. (2021). Prevention is better than cure: Machine learning approach to conflict prediction in Sub-Saharan Africa. *Sustainability 13* (13), 7366. https://doi.org/10.3390/su13137366.

Ngo, R. (2020). *AGI safety from first principles*. AI Alignment Forum. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. *Proceedings of the 23rd International Conference on World Wide Web — WWW '14*, 677–686. https://doi.org/10.1145/2566486.2568012.

NSCAI. (2021). *2021 final report*. https://www.nscai.gov/2021-final-report/.

O'Brien, J. T., & Nelson, C. (2020). Assessing the risks posed by the convergence of artificial intelligence and biotechnology. *Health Security 18* (3), 219–227. https://doi.org/10.1089/hs.2019.0122.

Ord, T., Mercer, A., & Dannreuther, S. (2021). *Future proof: The opportunity to transform the UK's resilience to extreme risks*. https://www.longtermresilience.org/futureproof.

Ourmazd, A. (2020). Science in the age of machine learning. *Nature Reviews Physics 2* (7), 342–343. https://doi.org/10.1038/s42254-020-0191-7.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. *ArXiv:1602.02697 [Cs]*. http://arxiv.org/abs/1602.02697.

Patty, J. W., & Penn, E. M. (2014). *Social choice and legitimacy: The possibilities of impossibility*. Cambridge University Press. https://doi.org/10.1017/CBO9781139030885.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today 26* (1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010.

Radford, A., Wu, J., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better language models and their implications. *OpenAI*. https://openai.com/blog/better-language-models/.

Raji, I. D., & Buolamwini, J. (2019). *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products*. MIT Media Lab. https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). Auditing radicalization pathways on youtube. *ArXiv:1908.08313 [Cs]*. http://arxiv.org/abs/1908.08313.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., . . . Bengio, Y. (2019). Tackling climate change with machine learning. *ArXiv:1906.05433 [Cs, Stat]*. http://arxiv.org/abs/1906.05433.

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schwalbe, N., & Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet 395* (10236), 1579–1586. https://doi.org/10.1016/S0140-6736(20)30226-9.

Seger, E. (2021). *The greatest security threat of the post-truth age*. BBC Future. https://www.bbc.com/future/article/20210209-the-greatest-security-threat-of-the-post-truth-age.

Seger, E., Avin, S., Pearson, G., Briers, M., ÓhÉigeartaigh, S., & Helena, B. (2020). *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*. https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf.

Smith, L., Peach, K., Ramos, J., & Sweeney, J. A. (2019). *Our futures: By the people, for the people*. https://www.nesta.org.uk/report/our-futures-people-people/.

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv:2102.02503 [Cs]*. http://arxiv.org/abs/2102.02503.

Toews, R. (2020). These are the startups applying AI to transform healthcare. *Forbes*. https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/.

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY 35* (1), 147–163. https://doi.org/10.1007/s00146-018-0845-5.

UN Security Council. (2021). *Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council*. https://digitallibrary.un.org/record/3905159?ln=en.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race and power in AI*. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf.

Whittlestone, J., Arulkumaran, K., & Crosby, M. (2021). The societal implications of deep reinforcement learning. *Journal of Artificial Intelligence Research 70* , 1003–1030. https://doi.org/10.1613/jair.1.12360.

Whittlestone, J., & Clark, J. (2021). Why and how governments should monitor AI development. *ArXiv:2108.12427 [Cs]*. http://arxiv.org/abs/2108.12427.

Whittlestone, J., & Ovadya, A. (2020). The tension between openness and prudence in AI research. *ArXiv:1910.01170 [Cs]*. http://arxiv.org/abs/1910.01170.

Wu, J., Leung, K., & Leung, G. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *The Lancet 395* (10225), 689–697. https://doi.org/10.1016/S0140-6736(20)30260-9.

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., Beaty, K. A., Dehan, E., & Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: Applications, challenges and future perspectives. *Human Genetics 138* (2), 109–124. https://doi.org/10.1007/s00439-019-01970-5.

Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., & Aliper, A. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews 49* , 49–66. https://doi.org/10.1016/j.arr.2018.11.003.

Zwetsloot, R., & Dafoe, A. (2019, February 11). *Thinking about risks from AI: Accidents, misuse and structure*. Lawfare. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

CHAPTER 3

ALIGNED WITH WHOM?

Direct and Social Goals for AI Systems

ANTON KORINEK AND AVITAL BALWIT

Introduction

We are building artificially intelligent systems which are increasingly competent and general. Already, AI systems can have impacts that occur almost instantly and anywhere in the world.[1] While speed and reach are factors that AI shares with other technologies, like malware or nuclear weapons, AI systems are also different, becoming increasingly autonomous and behaving like agents of their own.

When we delegate tasks to AI systems, an important challenge is to endow the systems with "desirable goals"—this is frequently labeled the AI alignment problem. (For a good survey of recent research on AI alignment see Ngo [2020]). However, when speaking of alignment, it is important to be explicit about whose goals with which an AI system is aligned. This chapter distinguishes between two types of AI alignment that both play critical roles in AI development, but that require distinct governance approaches:

> **Direct alignment:** when an AI system is pursuing goals consistent with the goals of its operator, irrespective of whether it imposes externalities on other parties.[2]
>
> **Social alignment:** when an AI system is pursuing goals that are consistent with the broader goals of society, taking into account the welfare of everybody who is impacted by the system.

There are multiple challenges to achieving direct alignment, including identifying the right goal to give to an AI system, conveying that goal, and getting the AI system to correctly implement the goal.[3] Social alignment adds the challenge of considering the interests of all of those impacted—not merely the operator of a system—when determining what goal to give the system.

What would it entail to consider all those affected? We describe a conceptual benchmark for social alignment based on welfare economics that we call *ideal social alignment* and analyze how social alignment differs from direct alignment in that it considers externalities. We provide examples of ideal social welfare functions that could, in principle, be employed to formally study social alignment. Because ideal social alignment is unattainable in practice,

we explain how it is possible to determine a partial ordering of social preferences, and that in practice, social alignment means aligning AI systems with this more limited set of instructions. Social alignment will generally require external interventions and a broader governance framework. We discuss how regulatory solutions provided by social norms, laws, markets, and architecture could be used to achieve social alignment, and we highlight the importance of AI governance to achieve social alignment.

We believe that ensuring AI systems satisfy both direct and social alignment will become ever more important in the future. Currently, we delegate fairly narrow tasks to AI systems—we ask them to classify images, rank our search results, recommend movies and music, and autocomplete bits of our emails. But we are starting to use AI in increasingly higher stakes environments, like evaluating loan and job applicants, financial trading, utility management, and national defense. Importantly, the range and complexity of tasks that we turn over to our AI systems are continuing to grow.

AI alignment—the challenge of how to endow AI systems with goals that are consistent with our goals—becomes ever more important as delegation from humans to AI systems (1) happens more often, creating more opportunities for misalignment to cause harm; (2) involves higher stakes situations, where misalignment would be more costly or even catastrophic; and (3) occurs in situations where we have less ability to provide oversight, which makes it more difficult to assess whether the system is aligned and less likely that we catch alignment failures early.

Concepts

We start by clarifying several theoretical concepts that are relevant for our discussion of alignment. Readers who are most interested in the comparison of direct and social alignment may wish to skip to the following section.

Agents and goals

We call entities that interact with their environment "agents" whenever it is useful to describe them as pursuing goals.[4] A goal is a summary description of what an entity is attempting to achieve. Humans clearly fit our description of agents—we think of ourselves and each other as pursuing goals, and this is useful because inferring goals makes for a more efficient description of what to expect next from the humans we interact with. For example, if a driver sees a person walking straight up to a crosswalk, it is useful to infer that the person's goal is to cross the street, and this is more efficient than to ponder how the person's leg movements will translate into the person's physical location over the ensuing seconds.

But our concept of agency is broader than humans. It also includes non-human entities, such as organizations or governments, which can be described as pursuing their own sets of goals. For example, a business organization may be described as following the goal of producing a product to earn profits; a university as advancing research and education; a government as pursuing the safety and well-being of its citizens.

And, importantly for the purposes of this chapter, our concept also includes artificial intelligence. Many modern AI systems are directly programmed to maximize a specific objective function, making them act in a goal-oriented way. More broadly, AI systems are agents in the sense that we define because they are designed to pursue specific goals, such as classifying images, driving cars, controlling robots, etc.

In our characterization of agency, we explicitly ask whether it is "useful" to describe an entity as pursuing a goal. This implies that the delineation is fluid and depends on the context. For example, if we start with a very simple mechanical structure and transform it into a progressively more intelligent robot, there is no specific threshold at which it becomes an agent; however, it will become more and more efficient and useful to describe it according to the goals it pursues rather than by the physical laws describing its mechanical structure. To provide further examples, it will rarely be useful to describe a rock as an agent, but there will be many situations in which it is useful to describe a Boston Dynamics Spot robot as an agent. Moreover, the same entity may be usefully described as an agent in some situations and contexts but not in others. For example, we may want to describe a robot as an agent while it is operating but as a piece of metal when we recycle it for scrap.

In the social sciences, an entity's "goals" (or objectives) are frequently described using a set of preferences; that is, an ordering that describes how the entity values different outcomes relative to each other. For example, a preference relation such as $A > B$ reflects that the entity prefers outcome A over outcome B.[5]

There is a dualism between an agent's goals and her actions—we can either describe the actions which emanate from her goals or describe the goals which drive her actions. When an agent's goals are fully specified in a given environment, the agent's actions are also fully specified (except where there are ties) because we can work out what actions the agent will find optimal to take, and vice versa, because we can infer what their goals are from their actions. We can simply translate back and forth from goal space into action space.

Simple depictions of agents in the social sciences frequently take an agent's goals as a primitive that is exogenously given (e.g., by postulating a certain utility function for the agent). Then they proceed to analyze what actions the agent would take to achieve the most desired outcome. However, such descriptions risk over-simplifying the behavior of human agents by leaving out subgoals, competing goals, or constraints on the described goal. For example, humans track the distance between current and desired states across multiple value dimensions that are in tension with each other in ways that change dynamically over time (see, for example, Juechems & Summerfield, 2019). This may pose significant challenges in determining human goals and may, for example, produce behavior that is inconsistent with any utility function.[6] As we will discuss in more detail, the difficulty of correctly determining goals is a major challenge for AI alignment.

Delegation and alignment

Delegation is when an entity charges another with the fulfillment of her goals. Following the language of economics, we use the term "principal" for the entity that delegates a task and "agent" for the entity who is charged with the task.[7]

Examples of delegation are as old as humanity itself and could already be found in hunter-gatherer societies.[8] As societies became more complex and hierarchical, the delegation of tasks and the need for goal alignment also started to involve governmental or religious organizations and, later, corporations. Humans and organizations both started to delegate tasks to each other to better accomplish their goals. For example, entrepreneurs founded corporations to pursue their goals, and corporations hired workers.

Successful delegation is advantageous for the principal by enabling her to better accomplish her goals. The reasons for this advantage include that the agent may have a comparative advantage in the task at hand, such as different or greater capabilities than the principal, better knowledge, or simply a lower opportunity cost of time. To make delegation successful, there needs to be a sufficient degree of alignment between the principal and the agent, but the principal also needs to provide the agent with a sufficient degree of discretion—a point that is not typically emphasized in traditional principal–agent theory in economics but that was already made by Weber (1922), and that is also emphasized in the recent literature on AI governance (see, for example, Young et al., 2021). In particular, the agent needs sufficient freedom of action to employ her greater capabilities, her better knowledge and judgment, or her additional time to be useful to the principal. However, this freedom of action is what creates problems when there is misalignment.

Economists and other social scientists have long studied principal–agent relationships; that is, how to align the actions of agents with the outcomes desired by principals (see, for example, Jensen & Meckling, 1976, for one of the most influential contributions). In that body of work, the principal and agent have exogenously given goals that differ from each other, and imperfect information makes it difficult for the principal to observe whether the agent has acted in her best interest or has abused the discretion she was afforded. The main research question centers around how the principal can provide the agent with incentives to act in her interest; for example, how to include carrots and sticks in work contracts to incentivize workers to exert the optimal level of effort.

In AI alignment, by contrast, the question is different and, in some sense, more fundamental: how to endow an agent with goals that lead to outcomes that are desired by the principal.[9] AI alignment is usually described as goal alignment, but what ultimately matters for the principal are the agent's actions. When the relationship between goals and actions is clear and is known perfectly, and when the goals of the principal and agent coincide, then there is no direct alignment problem. In Figure 3.1, this would correspond to each of the mappings that are indicated with arrows holding perfectly; that is, (1) the principal has identified the goals that will lead to her desired actions well, (2) they are correctly conveyed to the agent, who (3) in turn translates the conveyed goals into the desired actions. However, this is an idealized benchmark.

Delegation	Principal		Agent
Action space	Desired actions		Pursued actions
	↕(1)		↕(3)
Goal space	Desired goals	↔ (2)	Pursued goals

FIGURE 3.1 Principal–agent alignment in action space and in goal space

In practice, misalignment between the principal's desired and realized outcomes can arise in any of the three steps outlined in the figure. This allows us to distinguish the source of alignment problems into the following categories:

(1) **Identifying the principal's desired goals.** The principal needs to figure out what her goals are.
(2) **Conveying the goals to the agent.** Next, the principal needs to correctly transmit her desired goals to the agent so the agent can pursue them.
(3) **Translating the goals into actions.** The agent needs to correctly implement the transmitted goals by pursuing the corresponding actions.

We will elaborate on each of these steps in detail.

As we observed before, the term "AI alignment problem" is frequently used to describe step (2), i.e., how to provide an AI system with a set of goals that correctly reflect our goals, whereas the term "AI control problem" is used to capture the broader challenge of how to ensure that the actions of an AI system are desirable.[10] However, the three steps illustrated in our figure are closely related to each other. If the mappings described in steps (1) and (3) held perfectly, it would be possible to focus exclusively on how to align the goals of human principals and AI agents. In practice, however, it will be necessary to consider all three steps simultaneously as they cannot be cleanly separated from each other.

Direct and social alignment of AI

So far, we have discussed how to ensure that the actions performed by an AI system reflect the desires of a principal. However, we have been silent on who exactly the principal is. We distinguish two separate concepts, direct alignment and social alignment, that relate to whether we view the principal as the operator of an AI system or as society at large. The distinction between the two has not been sufficiently recognized in the existing literature on AI alignment.

As defined in the introduction, we use the term *direct alignment* to refer to whether an AI system is pursuing goals that are consistent with the goals of its operator, and *social alignment* to refer to whether an AI system is pursuing goals that are consistent with the broader goals of society, taking into account everybody who is affected by the system and internalizing any externalities. The two forms of alignment problems are also related to the broader challenge of developing cooperative AI (see, for example, Dafoe et al., 2020).

We will discuss the challenges of direct and social alignment in detail and will elaborate further on when the two differ from each other. But before doing so, let us provide two examples to highlight the difference between the two.

Example 1: Ted develops a new résumé screening algorithm. Because he wants the system to be free of racial bias, he leaves out the variable "race" from the training dataset. However, the system quickly learns correlates of race, such as name, address, and educational institution from the existing bias in the training dataset and uses these to arrive at biased hiring decisions.

This represents a failure of direct alignment. Ted was eager to avoid racial bias but did not realize that his implementation led to precisely the bias that he was concerned about.

Example 2: Mark develops a recommendation model to maximize user engagement on a social network platform. When he finds out that the system leads to stark increases in political polarization, he does not change course. This represents a failure of social alignment. The AI system pursued—and successfully achieved—its assigned goal, but it imposed large externalities on society by increasing polarization.

Direct Alignment

From the perspective of a principal operating an AI system, there are three interrelated challenges to ensuring direct alignment of an AI system: (1) the challenge of determining what goals to pursue, (2) conveying the goal to an AI system, and (3) getting the AI system to correctly translate the goal into actions.

Determining the goal

The first challenge is to work out the principal's goals; that is, to translate the principal's desired outcomes into abstract goals. Determining what we want can be difficult. Human goals are not easily interpretable: they are often amorphous or intuitively understood but difficult to express. Our brains are opaque and pursue multiple value dimensions depending on circumstances (see, for example, Juechems & Summerfield, 2019). When we are not sure what we want, it is difficult to align an AI system with our goals.

A key aspect of determining the principal's goal is how to scope the goal appropriately so that it does not conflict with other goals that are valuable to the principal. Humans have a broad set of goals that involve many different subgoals which we automatically and often subconsciously weigh against each other when they are in conflict. When we determine what goals to assign to AI systems, we must ensure that the systems do not optimize one subgoal to the detriment of others. This becomes more and more important as our AI systems become more powerful and their capacity to optimize over a single goal increases.

Example 3: Jack develops a recommendation model to maximize user engagement on a social network platform. He is dismayed to find out that the system also increases political polarization. This is a classic example of specifying an excessively narrow goal and obtaining unexpected side effects. In the described example, Jack did not anticipate that his recommendation model would also affect the political views of its users.

Conveying the goal

After a principal determines the content of her goals, she faces the technical challenge of transmitting the goal to an AI system. When humans convey goals to each other in natural language, they understand the context, which makes it easier to resolve ambiguities.

It is more challenging to convey goals to AI systems. AI systems do not share the same understanding of the world that we humans share and will likely not be able to resolve ambiguities by making commonsense deductions. Instead, we need to translate the goal into something machine readable and provide the instructions, training, or feedback necessary for the system to "understand" and execute. Part of the challenge is to clarify what our concepts mean. For example, if we told an AI system to "make us happy," what do we mean by that term? Do we mean pure hedonic experience, do we mean general life satisfaction, or any other range of viable meanings?

Example 4: Tim tells his AI-powered smartphone assistant to "call Jon" as he gets ready to go out and party. He is embarrassed that his 11 pm phone call wakes up his boss rather than reaching his brother, who is listed as "Jonathan" in his contact list.

This is an example of an AI system misinterpreting the goal of its principal because it did not correctly understand the context. A human assistant would have known not to call a work contact late at night, and he or she would have understood that "Jon" may refer to "Jonathan."

Sometimes, the challenges of determining and conveying can be addressed jointly. For example, inverse reinforcement learning allows an AI system to learn the objective function of its principal through observing their behavior (Ng & Russell, 2000). Through this method, the content and form of the goal blend together.

Implementing the goal

Once the principal's goals are transmitted to the agent, they need to be implemented through appropriately chosen actions by the AI system. Implementation has many technical aspects that we will not discuss here.[11] These include ensuring a system is free of bugs and is robust, including that it operates reliably in circumstances other than what it was initially trained and tested on. While implementation is key for getting a result that the principal is happy with, some do not view it as a pure component of alignment.[12]

Example 1 above described a résumé screening system that resulted in biased hiring decisions because of biased training data that was not corrected for. This is an example of an implementation failure.

Distinguishing direct alignment from social alignment

The described challenges of direct alignment also apply to social alignment. The key difference is that the first step, determining the goal, no longer involves a single principal who is operating the system, but instead the broader goals of others in society who would be impacted by the AI system. This is what makes the social alignment of AI a central theme of AI governance.

Many contributions in the existing literature on AI alignment refer to either direct alignment or social alignment without explicitly addressing the distinction between the two and the necessity of paying attention to both. For example, Paul Christiano (2018b) appears to focus largely on direct alignment in his definition of intent alignment: "AI A is aligned with an operator H, if A is trying to do what H wants it to do." It is possible that the operator in

this definition wants the AI to do something that will impose large externalities on others. Yudkowsky (2004) seems to describe a form of social alignment in defining an alignment benchmark that he terms coherent extrapolated volition (CEV) as, "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together." He explicitly refers to "*our* wish . . . had [*we*] grown up farther *together*" (p. 6).

Other definitions are unclear as to whether they imply direct or social alignment. Christiano (2018a) defines the project of alignment as "aligning AI systems with human interests," leaving unspecified which humans these interests belong to. For example, Stuart Russell (2019, 149) defines the problem of value alignment as ensuring that we do not "perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own." Brian Christian (2020), author of *The Alignment Problem*, defines the titular problem as "how to ensure that these models capture our norms and values, understand what we mean or intend, and, above all, do what we want" (19). Both definitions do not make it clear who "we" or "our" refers to—all of humanity, some representative group, a synonym for one? Evan Hubinger (2020) writes that an AI agent is "impact aligned (with humans) if it doesn't take actions that we would judge to be bad/problematic/dangerous/catastrophic." It is not clear who these "humans" are—just the principals/operators, or all humans?—or how large a consensus this "we" must attain.

Social Alignment

In contrast to the direct alignment problem, which looks at whether an AI system is consistent with the goals of its operator, the social alignment problem looks at whether the system pursues goals that are consistent with the goals of society at large.

In the following, we start by describing a benchmark for social alignment based on welfare economics that we call *ideal social alignment*. We analyze how social alignment differs from direct alignment by emphasizing that it considers externalities on others. We provide examples of ideal social welfare functions that could, in principle, be employed to formally study social alignment. However, we also lay out the practical difficulties of establishing what society's preferences are, and we analyze how to deal with situations in which there is no clear way of mediating conflicting objectives within society. (See also Baum [2020] for a description of the difficulties in establishing society-wide ethics for AI.)

Next, we analyze to what extent existing norms (broadly defined) for the operators of AI systems are sufficient to ensure that the systems are socially aligned. We distinguish two sets of problems that give rise to misalignment: the first arises when the operators of AI systems violate existing social norms and are themselves not socially aligned; and the second when our current social norms are insufficient to address some novel externality caused by new AI capabilities. Achieving social AI alignment requires progress on both fronts. We next describe a set of policy tools that can be used to further the social alignment of AI.

Ideal social alignment

To describe a simple idealized benchmark for social AI alignment, we impose two assumptions. First, we assume that society has a complete and well-defined set of preferences over all choices

that are relevant for the AI system. Second, we assume that society itself creates, operates, and fully controls the AI system. Under these idealized conditions, the challenges of social alignment of AI would be reduced to the challenges of direct alignment that we described in the previous section because society would essentially act as the operator.

Technically speaking, the first assumption of complete preferences implies that society's preferences define a full ordering of all the available choices; that is, that whenever society faces a set of choices, say A, B or C, there are social preferences that tell us how they rank compared to each other, which ones are preferred, or which ones are equally desirable for society.[13] One possible ranking would be, for example, A > B > C. As we will explore further below, this is a strong assumption: in practice, there are many areas in which society does not have such clear preferences.

Theoretically, ideal social alignment could also be achieved if an AI system is operated by a true altruist who implements society's preferences. Most human beings are not purely egoistic but also care about the well-being of others—they have altruistic preferences, which means that they explicitly consider and accommodate the preferences of a broader group than just themselves. Their consideration of the welfare of others may range from minor (i.e., they care about others much less than about themselves) to perfect altruism (i.e., their decision making considers the welfare of another person just as much as their own). The direct alignment problem only considers the effects of the AI system on others to the extent that the person or entity creating the system is altruistic. If those who set AI goals are perfectly altruistic, like this theoretical altruist, then the social alignment problem is solved—their individual goals will be aligned to social goals. However, whenever the preferences that the altruist perceives about society differ from society's actual preferences, or whenever society's preferences change, the supposed altruist would become a dictator. This is a significant risk whenever there is no democratic process that represents a corrective force to express society's preferences.

As reflected in the term, *ideal social alignment* is more a conceptual benchmark than a guideline for actual implementation. Even in a well-functioning representative democracy, it is unlikely that society will agree on all outcomes. However, the benchmark is useful to sharply delineate the differences between direct and social alignment.

At the center of the distinction between direct and social alignment are *externalities*. In our context, externalities exist whenever an AI system affects others without their agreement and without the beneficiary compensating others for it. For example, an AI system that engages in surveillance of individuals or otherwise intrudes into their privacy imposes externalities onto them. Likewise, a system that manipulates consumers into buying goods on false pretexts inflicts externalities on its victims. Similarly, a system that evaluates applicants for jobs, loans, or other transactions but engages in unfair discrimination imposes externalities on its subject. At a society-wide level, a system that promises social engagement to its users but creates echo chambers and deteriorates democratic processes while doing so inflicts significant externalities on society at large. Similarly, AI systems that displace human jobs without generating sufficient new jobs depress economy-wide labor demand, reduce wages, and contribute to rising inequality. These effects represent what economists call pecuniary externalities because they arise from changes in market prices (in the given case, declines in workers' wages). AI systems may also confer positive externalities on society, such as a tool that aids in scientific discovery which may have benefits for society that are far greater than the benefits the developer of the tool obtains.

Framed this way, the goal of social alignment is to internalize the externalities of AI systems, making sure that AI systems consider the benefits and costs not only for their operators but also for all other members of society. In some cases, society may view the potential externalities of an AI system as so harmful that it is best for the system not to be implemented (e.g., many forms of mass surveillance may simply be too harmful to be worth any benefits they create). In general, however, it is not necessarily desirable to reduce all externalities to zero, just to properly take them into account. For example, an AI system that creates large amounts of value at the cost of displacing some jobs may well be worth implementing. Social alignment simply requires that the balance of the useful and deleterious effects reflects what society would choose as the optimal balance.

Social welfare functions are a widely used concept to weigh benefits and costs that can be applied to the ideal social alignment of AI systems. Social welfare functions are a generalization of utility functions to social choices and work in a similar fashion: given a set of possible choices for society, a social welfare function assigns to each choice a welfare score that captures how desirable it is for society, and the choice that obtains the highest score is by definition the preferred one that society would adopt. A widely used example are utilitarian (also called Benthamite) welfare functions, which sum up the utilities of all members of society with equal weight. A specific case among utilitarian social welfare functions that evaluate consumption choices is when individuals are attributed utility functions that are linear in consumption. In that case, social welfare is simply the sum of all consumption, and maximizing social welfare is equivalent to maximizing economic efficiency; that is, maximizing total consumption without any regard for distributive concerns. Conversely a Rawlsian (or maximin) social welfare function captures the desire to maximize the utility of the worst-off member of society, reflecting a strong desire for equality.

The analogy to utility functions makes social welfare functions an appealing concept for AI developers—utility maximization is a cornerstone of training AI systems. This makes it important to be aware of the limitations of social welfare functions. One limitation of social welfare functions as commonly used is that they focus on consequentialist specifications of individual utility, which may miss other considerations that society views as important for ethical decision making (e.g., deontological considerations).[14] Imposing a bad welfare function, even with the best of intentions, will create social harm. Another important limitation is that social welfare functions pre-suppose a complete set of social preferences that are well-defined over all choices, but there are many situations in which this may not be achievable. This is what we will turn to next.

Disagreement and partial social preferences

Social alignment can only occur where there are well-defined social preferences. However, members of society frequently disagree on what are the most desirable choices. Unfortunately, there is no general way to compile the preferences of multiple individuals into well-defined and rational social preferences over all available choices. Condorcet (1785) already observed that democratic voting will not in general produce a full set of rational social preferences.[15] Arrow (1950) showed in his doctoral thesis that this is a general property of all mechanisms to aggregate individual preferences into social preferences, except for a dictatorial rule whereby a single member of society dictates all choices.[16] These negative results imply that it

is generally not possible to come up with the complete set of preferences that would be necessary for AI systems to implement the ideal social alignment that we described. However, that does not imply that we need to give up on social alignment entirely.

Instead, social alignment can still focus on aligning AI systems with those social preferences that can be clearly established. Even though members of society may not agree on how to rank all available choices, they will agree on how to rank many of the most important choices. They may agree that A > B and A > C but may not be able to rank B and C relative to each other. The social choices over which there is general agreement within society represent a partial ordering of all the available choices. We can increase the set of choices for which a social preference can be established if we weaken the standard from universal agreement to somewhat lower standards, such as near-universal agreement. For example, society will generally agree that it is desirable to save lives, or to refrain from actively discriminating against minorities, even if a small fraction of the population disagrees. The resulting partial ordering provides a limited set of instructions for social choices, even though it cannot identify a full set of social preferences that apply to all circumstances.

Social alignment requires that AI systems observe the partial ordering provided by social preferences. Formally, we call an AI system socially aligned if its choices correspond to the partial ordering implied by social preferences. Conversely, an AI system violates social alignment if it makes choices that contradict the partial ordering implied by social preferences, that is, if society generally agrees that it would make different choices.

Given the partial nature of the ordering implied by social preferences, there will be situations in which society genuinely disagrees so social preferences do not provide instructions for what a socially aligned AI system should do. In those instances, existing social norms cannot provide guidance for the AI system or its operator about what choices to make. In some contexts, there may not be a preferred social choice, but there may be agreement among members of society that AI systems and their operators should have the liberty and freedom to make their own choices, such as the freedom of how to design a new product or how to compete in the market (while respecting the rules). In open societies, such freedoms are in themselves an important value.

However, in other contexts, unresolved conflicts within society imply that the choices of AI systems and their operators are likely to be contentious, no matter what choice they make. One of the fundamental goals of governance is to resolve such conflicts, to determine how to establish social preferences in such situations and how to establish norms that encapsulate these preferences. The question of what social preferences ought to be in the realm of AI—and over which social choices they are defined—is a recurring theme of this *Handbook* and, more generally, a central theme of AI governance.

Rights-based approaches are a common mechanism by which society represents partial social preferences because they reflect certain defined entitlements and freedoms for members of society while leaving ample space for other choices over which there may be disagreement. For example, one of the most fundamental rights-based approaches, human rights, encapsulates a set of basic rights and freedoms that all human beings are entitled to Universal Declaration of Human Rights (UN General Assembly, 1948) . Among the countries adopting the declaration, there is general agreement on these rights. In the digital realm, an example of a rights-based approach is the EU's General Data Protection Regulation European Commission (2016), which adds many new rights that have become relevant only recently, such as the right of access to information or the right to be forgotten.

Bajgar and Horenovsky (2021) describe how rights-based approaches may be useful for long-term AI safety and regulation.

Spheres of social alignment

The social alignment problem can manifest at several different scales, ranging from small subgroups of society to larger spheres such as humanity as a whole. The commonality between them is that they all include cases where someone other than the AI system's operator is affected by externalities.

In general, the extent of agreement within a group is decreasing with group size. Smaller communities may find it easier to come to an agreement on what outcomes are desirable for an AI system to pursue than larger groups of people, such as the citizens of a nation or humanity as a whole. This implies that group preferences will lead to increasingly partial orderings as the group size increases; that is, larger groups will agree less on what outcomes to pursue than smaller, more homogenous groups.

A good example of how attitudes towards social alignment differ depending on group size is when there are competitive dynamics between subgroups of society. Consider two large corporations developing AI systems that are in fierce competition with each other. When one corporation improves its system, it gains market share at the expense of the other. The stakeholders of each corporation, including its workers, shareholders, and suppliers, form a subgroup of society, and this subgroup has a clear interest in the corporation doing well. If a corporation's AI system pursues outcomes that are in the collective interest of that subgroup, then the system is socially aligned at the subgroup level; that is, at the level of the corporation.

Looking outside the corporation, there are clearly negative externalities between the two corporations. However, our social norms at the national and international level allow for such competition and do not find competitive dynamics objectionable as long as they benefit consumers and satisfy other applicable laws. The partial ordering reflecting our society-wide norms includes the requirement to act lawfully, to avoid biases, etc., but gives corporations freedom to engage in lots of actions, including the freedom to compete with each other. Under the described circumstances, the corporation's AI system is aligned at the society-wide level even though it imposes large negative externalities on competitors.

For purposes of illustration, and without being exhaustive, we discuss a few different exemplary spheres of social alignment.

Group-level

A social alignment problem could manifest at the group level, such as within a community, club, university, corporation, or city. If an AI system run by members of the group imposes externalities on the group that violate the social norms within the group, it is socially misaligned at the group level.

Country-Level

For many questions, individual countries are the most important sphere at which to consider social alignment. In the modern world (perhaps with the exception of the EU), most

laws and regulations originate at the country level because countries are the politically most powerful actors. This is also true for AI regulation to forestall alignment problems.

However, social alignment at the country level is not necessarily sufficient. Countries are frequently subject to competitive dynamics, especially in the military context, where advances in AI may give rise to significant shifts in power dynamics (see, for example, Armstrong et al., 2016). This directly leads to the next and widest sphere at which social alignment is desirable.

World-level

Social alignment at the world level is the broadest, least restrictive, but perhaps also most fundamental sphere of alignment for AI systems. It requires that an AI system pursues desired outcomes on which humanity at large broadly agrees. Although there are many areas of significant disagreements among the world's citizens, there are also areas of almost-universal agreement. Examples include the desirability of basic forms of AI safety to avoid human extinction, or that the most momentous decisions undertaken by autonomous weapons systems should have humans in the loop (Human Rights Watch, 2012). Another area of near-universal agreement may be that it is undesirable to develop a super-human AI system that displaces all human labor without ensuring that humans have sufficient material resources to survive such a radical shift. Articulating and formalizing global social norms on these topics is a pressing area of concern.

Implementing social alignment

No matter in which sphere, attaining social alignment of an AI system may be more challenging than attaining direct alignment. The main goal for the creator of an AI system is to solve the direct alignment problem so that the system pursues the outcomes he desires. Social alignment may be a second thought.

The social alignment of an AI system would be ensured if its operator is perfectly altruistic and internalizes all externalities that the system imposes on others in an ethical fashion. However, more generally, the operator may not care about imposing harm on others as long as the system achieves the desired outcomes. We will now discuss the available avenues to achieve social alignment.

Assessing AI Impacts.

A precondition for evaluating social alignment is knowing about the impacts and potential externalities generated by an AI system. Sometimes harms arise without the operator of an AI system even being aware of them, and the operator may not have sufficient incentives to find out. Moreover, lack of transparency makes it easier to cover up harms. AI impact assessments could help. Lessons can be learned from environmental impact assessments (EIAs), which are routinely required for actions of government agencies, or for government-funded, -permitted, or -licensed activities, such as building a highway, airport, or oil pipeline. AI impact assessments could quantify the potential risks and benefits of AI systems. Such assessments could be mandated for AI projects that are implemented by government

entities, that receive government funding, or that have sufficiently broad societal effects. They could also be used on a voluntary basis, just like EIAs have become relatively common in the private sector.

Existing Norms

Once the potential externalities of an AI system are known, the next question is how social norms, regulations, and laws can ensure the social alignment of AI systems. Society already has a rich set of norms for social alignment that have evolved over centuries, consisting of informal social customs and habits, as well as formal laws and regulations. These norms constrain the behaviors of both individual humans and non-human agents, such as governments, corporations, or nonprofits. AI systems operated by these agents are by extension also subject to those norms. To provide a stark example, a civilian must not program an AI-based robot to kill someone.

The social alignment of the operators of AI systems thus leads to a certain "default" level of social alignment for AI. Conversely, when these actors violate the social norms that they are subject to, they give rise to alignment problems that we may call social misalignment from violating existing norms. We start by discussing how informal and formal social norms address social alignment. Then we discuss why we believe that it is also increasingly desirable to impose new constraints directly on AI systems in addition to existing norms on their operators to guarantee that AI systems are socially aligned.

Informal Social Norms

Informal social norms are constraints on agents' behavior that are enforced in an informal, decentralized way by a community. They have evolved together with humanity to facilitate human cooperation. They can be seen in action, for example, when employees, consumers, or shareholders pressure companies to abstain from behaviors that they view as unethical.

The importance of informal social norms is frequently underemphasized, such as in economic analyses when individuals are counterfactually depicted as perfectly selfish actors. There is significant room for improving social AI alignment by establishing the right informal norms within the AI ecosystem (see, for example, Klinova, 2024). Such norms can be powerful in driving the behavior of individual humans. For example, informal social norms among AI developers as to what types of systems are considered ethical and desirable and what is considered unethical provide effective constraints on what systems AI companies develop. We can already see the effect of such norms on AI development in that many AI companies have begun creating codes of ethics or employing teams that directly focus on AI ethics and society (Bessen et al., 2021). Similarly, informal social norms among the broader public can translate into consumer pressure, such as not purchasing from AI companies that don't live up to their expectation of social alignment.

However, informal social norms alone are insufficient to govern our complex modern societies. They are most effective at the human community level. Non-human entities such as corporations and governments are not directly susceptible to informal social norms. Instead, they are only susceptible indirectly via their human agents, which opens the door

to what some have called administrative evil (see, for example, Young et al., 2021). More formal governance modes, such as laws and regulations, are therefore indispensable.

Laws and Regulations

Laws and regulations impose constraints on agents that are enforced with formal, state-backed or -administered penalties. There are several ways in which such legal constraints can contribute to social alignment.

- Prohibitions and mandates. There are some uses of AI that society will deem too harmful to allow, and for these it may make sense to pass legislation which forbids them. For example, the Campaign to Stop Killer Robots has launched an effort to ban lethal autonomous weapons that could kill without human oversight.[17] Similarly, mandates can be enacted to ensure that AI systems meet certain socially desirable minimum standards, such as in the realm of safety. A closely related measure is to assign harmed individuals rights that can be enforced via litigation (see, for example, Kessler, 2010).
- Taxes and subsidies. As a classic fix for externalities, they apply just as well to the case of social AI alignment. They are preferable to outright bans and mandates when an activity creates externalities so the unregulated amount of that activity would be undesirable, but when a total ban would be excessive. Moreover, the revenue from taxes can be distributed to those who experience the harms.[18] For example, tracking of individuals represents a privacy intrusion but may also offer some useful benefits. Instead of banning it outright, imposing taxes or user fees may reduce it to a more desirable level. Similarly, if society's goals include an equitable income distribution, excessive automation that destroys jobs and undermines worker incomes could be taxed, with the resulting revenue distributed to workers losing their jobs, but it would be undesirable to ban automation.

New norms for social AI alignment

The growing powers and capabilities of AI systems are creating new and ever more powerful ways in which the public interest may be infringed upon (new externalities), which call for new social norms and create new potential forms of social misalignment. When AI systems gain new capacities, society can be unprepared to govern them. For example, society had one set of norms for surveillance and privacy when surveillance was very labor-intensive and correspondingly costly to undertake. As a result, governments focused surveillance only on very high-value targets, such as suspected high-value criminals. Legal constraints on mass surveillance would have been redundant, given the high cost of surveillance. Now that AI systems can perform many forms of surveillance cheaply at large scale, new legal constraints on surveillance activities have become necessary.

More generally, every time a new AI capability is developed, it may bring up new social alignment problems that call for new social norms. Aside from privacy norms, additional examples in areas that we already touched upon include the need for new norms

for AI systems that become increasingly adept at manipulating consumers. Moreover, we need new fairness norms for AI systems that make high-impact decisions that have hitherto been reserved to humans. As the labor market effect of AI and other forms of automation become more severe—and more pernicious for workers—there is also a new need for norms for when and how AI developers should compensate the exposed workers. Even more starkly, if AI makes human workers economically redundant, society will need to establish new norms for how to provide humans with income when labor income is no longer an option (see Korinek & Juelfs, 2024).

Social Alignment Norms Imposed on Whom?

Up until the recent past, governance to ensure the social alignment of AI systems has relied entirely on society imposing norms on the operators of AI systems, who would be charged with ensuring the alignment of their systems. This is the case, which we may call social alignment by extension, illustrated in panel (a) of Figure 3.2, which employs arrows to indicate that an entity imposes norms on another entity. Such an arrangement would be all that is needed if (1) the operator were perfectly aligned with the social norms, and (2) if the direct alignment between the operator and the AI system held perfectly.

However, when one of these two conditions is violated, it makes it desirable for society to directly impose social norms on AI systems, as illustrated in panel (b) of Figure 3.2. Let us consider each of the two conditions in turn.

When the operator of an AI system is not in compliance with social norms, then imposing social norms directly on AI systems may substitute for the operator's lack of compliance. Such an arrangement may also make it easier to monitor the social alignment of the operator. Consider, for example, an unethical corporation that pursues blind profit maximization to the detriment of other values of society. If the AI systems deployed by such a corporation need to satisfy certain enforceable norms, such as being unbiased, then the space for unethical behavior of the corporation is curtailed. In fact, norms imposed on AI systems may even make it possible to regulate behaviors that violate social norms but were difficult to regulate before. For example, when lending decisions were made individually by loan officers, it was harder to establish whether they were unbiased than it is with algorithms.

When an operator is generally aligned with social norms but has not fully solved the direct alignment problem between her and an AI system, then norms that are imposed directly on the AI system may also help. Such norms can be thought of as best practices, and

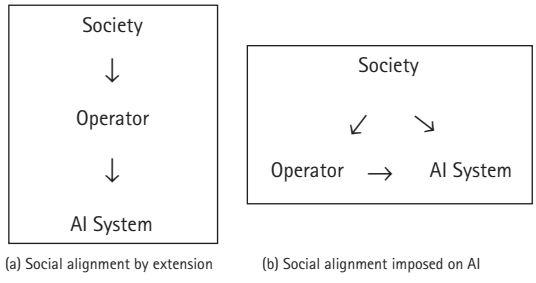

FIGURE 3.2 Two modes of imposing social alignment norms on AI systems

they may contribute to all three steps of the direct alignment problem that we previously explored: determining the right goal, conveying the goal, and implementing the goal. For example, they may help a well-intended but inexperienced entrepreneur to ensure that the AI system she develops does not unintentionally impose harm on society.

As AI systems become more agentic and have ever more discretion over decisions that used to be reserved for humans, we believe that imposing norms directly on AI systems is becoming increasingly important.[19]

Conclusion

As AI systems become more powerful and are deployed in a growing number of areas, aligning them with our goals becomes ever more vital. However, the expression "our goals" is often used too loosely. It is crucial to emphasize that AI alignment has two distinct dimensions: direct alignment and social alignment. The two dimensions require somewhat different approaches, but we need to solve both to ensure a future that is desirable for humanity. Direct alignment ensures that AI systems pursue goals consistent with the objectives of their operators, irrespective of whether they impose externalities on other parties. By contrast, social alignment ensures that AI systems pursue goals that are consistent with the broader objectives of society, internalizing externalities and considering the welfare of everybody who is impacted by them.

Modern AI systems have the capacity to powerfully optimize for the goals with which we endow them. They are becoming better and better at doing what we are asking them to do and reaching their programmed goals, no matter if these goals correspond to our true goals or if we mistakenly assign them the wrong goals, such as excessively narrow subgoals that lead to disastrous unintended side effects because they fail to fully capture what we want.

Regarding direct alignment, we need to work on determining, conveying, and implementing the goals that we want AI systems to pursue in a robust manner. Regarding social alignment, society needs to determine what social goals and norms we want AI systems to pursue. Social preferences can only determine a partial ordering over all available choices. It is important to expand that ordering as much as possible by resolving social disagreements and conflicts, and to appeal to our better angels as we do this so that our preferences reflect our ethical values. To the extent that society finds agreement, it is also important to develop the right institutions to implement our preferences. We argue that this requires imposing norms on the developers and operators of AI systems as well as new norms that are directly imposed on AI systems.

As AI systems have become more powerful and their use in our world has become more widespread in recent years, we have also witnessed a growing number of cases of social alignment failures, from automated decision systems with biases against disadvantaged groups to social networks that increase polarization and undermine our political systems. Yet progress is continuing, and the powers of our AI systems are continuing to evolve. This makes it urgent to accelerate our efforts to better address the social alignment of AI. If we already have difficulty addressing the AI alignment problems we face now, how can we hope to do so in the future when the powers of our AI systems have advanced by another order

of magnitude? Creating the right governance institutions to address the social AI alignment problem is therefore one of the most urgent challenges of our time.

Acknowledgments

We would like to thank Ondřej Bajgar, Damon Binder, Justin Bullock, Alexis Carlier, Carla Zoe Cremer, Allan Dafoe, Ben Garfinkel, Lewis Hammond, Fin Moorhouse, Luca Righetti, Toby Shevlane, Joseph Stiglitz, and participants at the Spring 2021 Handbook of AI Governance conference for helpful comments and discussions. Any remaining misalignment is our own. We gratefully acknowledge financial support from Center for Innovation, Growth and Society at the Institute for New Economic Thinking (CIGS-INET).

Notes

1. An example is AI-powered financial trading systems (see Boukherouaa & Shabsigh, 2021).
2. Throughout the paper, we use the convention of referring to the entity that is creating, operating, and controlling an AI system as the "operator." In principle, each of these tasks could be performed by different entities, adding additional complexity to the challenge of AI alignment.
3. There are many alternative ways of defining alignment but with similar flavor. For example, an AI system could be aligned to the human's instructions, intentions, revealed preferences, informed preferences, interests, or values, among other options. See Gabriel (2020) for a fuller discussion.
4. This is a shallow definition of agency that is, however, useful for our purposes here. It is inspired by, but distinct from, Dennett's work on stances (see Dennett, 1987). In different contexts, other definitions may be more useful. For example, in ethics, a moral agent is an entity that is morally accountable for its actions. For an elaboration on alternative concepts of agency, see Franklin and Graesser (1996) or Orseau et al. (2018).
5. Another way to express goals is in the form of a "utility function" u(X) that assigns a numerical value to each possible X which ranks the different possibilities. Utility functions are more restrictive than preference orderings. In other words, every utility function defines a set of preferences, but not every set of preferences can be captured by a utility function. For example, for lexicographic preferences, doing so is impossible.
6. Specifically, for a set of preferences to map into a utility function requires several technical assumptions that may be violated, including completeness, transitivity, and continuity.
7. Unfortunately, this nomenclature involves some overloading of the term "agent." In the previous section, we called any entity that can be described as pursuing a goal as an agent; in this section, we follow the conventions of principal-agent theory. Throughout the remainder of this article, the meaning of the term will be clear from the context.
8. In fact, some of the interactions among other social species such as bees or ants can also be described as simple forms of delegation.
9. On the surface, the two described situations—incentivizing a human agent with distinct goals to pursue the principal's goals versus creating an AI agent from scratch who pursues the principal's goals—seem very different. However, given the dualism between actions and goals, there is in fact a deeper equivalence between the two. Addressing the classic

principal–agent problem in economics can be viewed as a situation in which the principal has only limited ability to affect the agent's architecture (e.g., to reprogram the agent's primal drive to avoid hard work) and needs to find workarounds ("incentives") to make the agent pursue the desired goal. Programmers frequently experience similar situations. For example, the architecture of ML libraries, say TensorFlow, constrains how they can write their code and makes some results far easier to obtain than others. In other situations, they need to write workarounds building on clunky legacy applications to efficiently obtain the desired behavior. Conversely, human principals sometimes have the ability to "reprogram" agents. For example, parents greatly appreciate the importance of instilling proper goals into their offspring; managers and military leaders know the importance of "inspiring" their agents to pursue desired goals; and a significant part of our human culture (religion, morals, etc.) revolves around reprogramming humans' goals in a way to make our societies operate more harmoniously.

10. For example, Bostrom (2014) describes value alignment as one element of AI control alongside other mechanisms such as capability controls.
11. For a thorough and cutting-edge technical introduction see Russell and Norvig (2020).
12. For example, some definitions of alignment only capture the intentions of the AI system, not the outcome, whether ex ante the AI system was trying to achieve the human's goal. If an AI system tries to accomplish the goal, but some implementation failure causes the system to crash before doing so, perhaps it should not be viewed as a failure of alignment.
13. In mathematics, a full ordering (or total order) is a binary relation on a set that satisfies, among other conditions, that it is transitive and that any two elements are comparable (see https://en.wikipedia.org/wiki/Total_order). The assumption that society's preferences represent a full ordering is also a necessary condition for describing them via social welfare functions.
14. Although it is possible to add such considerations with a negative weight in consequentialist specifications of welfare functions, it is difficult to determine desirable weights.
15. Specifically, when three or more people are asked to express their preferences over three or more alternative choices in pairwise votes, they frequently arrive at outcomes like A > B, B > C and C > A, making it impossible to establish a full order of the available social choices.
16. Even seemingly straightforward mechanisms such as a welfare function that is the sum of utility functions of all members of society will not satisfactorily address the problem, since it may lead to Pareto-dominated outcomes. See Eckersley (2019) for a fuller description of the problem in the context of AI alignment.
17. See https://www.stopkillerrobots.org/.
18. Allowing the harmed individuals themselves to impose a user fee is equivalent to taxing the harm and distributing the revenue to the harmed individuals.
19. In related work, Korinek (2021) proposes the establishment of an AI Control Council to further these objectives.

References

Armstrong, Stuart, Bostrom, Nick, & Shulman, Carl. (2016), Racing to the precipice: A model of artificial intelligence development. *AI & Society 31*(2), 201–206.

Arrow, Kenneth J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy 58*(4), 328–346.

Bajgar, Ondrej, & Horenovsky, Jan. (2021). Human rights as a basis for long-term AI safety and regulation. Working Paper, University of Oxford.

Baum, Seth D. (2020). Social choice ethics in artificial intelligence. *AI & Society 35*(1), 165–176.

Bessen, James, Impink, Stephen M., Reichensperger, Lydia, & Seamans, Robert. (2021). Ethics and AI startups. *Scholarly Commons at Boston University School of Law*, July 30. https://scholarship.law.bu.edu/faculty_scholarship/1188.

Bostrom, Nick. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Boukherouaa, El Bachir, & Shabsigh, Ghiath. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance. Departmental Paper DP/2021/024, International Monetary Fund.

Christian, Brian. (2020). *The alignment problem*. W.W. Norton.

Christiano, Paul. (2018a). About AI alignment. *AI Alignment*. https://ai-alignment.com/about.

Christiano, Paul. (2018b). Clarifying "AI alignment." *AI Alignment Forum*. https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment.

Dafoe, Allan, Hughes, Edward, Bachrach, Yoram, Collins, Tantum, McKee, Kevin R., Leibo, Joel Z., Larson, Kate, & Graepel, Thore. (2020). Open problems in cooperative AI. Technical Report, DeepMind.

Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.

Eckersley, Peter. (2019). Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). *Proceedings of the AAAI Workshop on Artificial Intelligence Safety*, 1–8. https://dblp.org/rec/conf/aaai/Eckersley19.html?view=bibtex.

European Commission (2016), General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, available at: https://gdpr-info.eu/

Franklin, Stan, & Graesser, Art. (1997). Is it an agent, or just a program? A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings (Eds.), Intelligent agents III agent theories, architectures, and languages. ATAL 1996. Lecture Notes in Computer Science, vol 1193. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0013570.

Gabriel, Iason. (2020). Artificial intelligence, values, and alignment. *Minds & Machines 30*, 411–437. https://link.springer.com/content/pdf/10.1007/s11023-020-09539-2.pdf.

Hubinger, Evan. (2020). Clarifying inner alignment terminology. *AI Alignment Forum*. https://www.alignmentforum.org/posts/SzecSPYxqRa5GCaSF/clarifying-inner-alignment-terminology.

Human Rights Watch. (2012). Losing humanity: The case against Killer Robots., Technical Report. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

Jensen, Michael C., & Meckling, William H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics 3*(4), 305–360.

Juechems, Keno, & Summerfield, Christopher. (2019). Where does value come from? *Trends in Cognitive Sciences 23*(10), 836–850.

Kessler, Daniel P. (2010). *Regulation vs. litigation: Perspectives from economics and law*. University of Chicago Press.

Klinova, Katya. (2024). Governing AI to advance shared prosperity. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Korinek, Anton. (2021). Why we need a new agency to regulate advanced artificial intelligence: Lessons on AI control from the Facebook Files. Report, Brookings Institution, December 8. https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/.

Korinek, Anton, & Juelfs, Megan. (2024). Preparing for the (non-existent?) future of work. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Marquis de Condorcet (1785), Essay on the Application of Analysis to the Probability of Majority Decisions, Paris: Imprimerie Royale.

Ng, Andrew Y., & Russell, Stuart J. (2000) Algorithms for inverse reinforcement learning. *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, 663–670.

Ngo, Richard. (2020). AGI safety from first principles. *AI Alignment Forum*. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.

Orseau, Laurent, McGill, Simon McGregor, & Legg, Shane. (2018). Agents and devices: A relative definition of agency. arXiv. https://arxiv.org/abs/1805.12387.

Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, Stuart J., & Norvig, Peter. (2020). *Artificial intelligence: A modern approach*. 4th US edition. Pearson.

UN General Assembly, Universal Declaration of Human Rights, 10 December 1948, Resolution 217 A (III), available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights

Weber, Max. (1922). Bureaucracy (E. Fischoff, Trans.). Translation of Chapter 6 in *Wirtschaft und Gesellschaft* (pp. 956–1005). Mohr.

Young, Matthew M., Himmelreich, Johannes, Bullock, Justin B., & Kim, Kyoung-Cheol. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4*(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.

Yudkowsky, Eliezer. (2004). Coherent extrapolated volition. Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf.

CHAPTER 4

THE IMPACT OF ARTIFICIAL INTELLIGENCE

A Historical Perspective

BEN GARFINKEL

Introduction

Over the next several decades, artificial intelligence (AI) is likely to change the world in numerous ways. When trying to describe just how significant these changes could be, commentators tend to reach for historical comparisons. One influential researcher, Andrew Ng, has famously called AI "the new electricity" (Ng, 2017). Elsewhere it is possible to find analogies to fire (Clifford, 2018), nuclear weapons (Allen & Chan, 2017), industrialization (Brynjolfsson & McAfee, 2014), the first computer software (Karpathy, 2021), and even, on occasion, life itself (Tegmark, 2017).

This chapter also takes a history-oriented approach to discussing the impact of AI. Rather than drawing comparisons to individual technologies, however, the chapter instead situates AI within two "reference classes" of technologies that share common traits. If there are any common patterns in how technologies within these reference classes have impacted the world, then we might expect AI to display some of the same patterns.

The first reference class I consider is the set of *general purpose technologies* (GPTs). General purpose technologies are distinguished by their unusually pervasive use, their tendency to spawn complementary innovations, and their large inherent potential for technical improvement. Modern examples include computers, the internal combustion engine, and—in keeping with Ng's suggestion—electricity. Many economists now regard artificial intelligence as an emerging GPT.

I report a few key lessons from the literature on general purpose technologies. One lesson is that the early applications and iterations of GPTs tend to be unassuming. It normally takes several decades for them to achieve large-scale impacts. However, in the long run, a GPT can be expected to alter everything from economic productivity to the character of war to how people spend their leisure time. I attempt to apply these lessons to the specific case of artificial intelligence.

The other, strictly smaller reference class I then consider is the set of *revolutionary technologies*. A revolutionary technology is a GPT that supports an especially fundamental transformation in the nature of economic production. There are only two obvious examples of revolutionary technologies. The first example is domesticated crops, which supported the transition from hunting and gathering to widespread agricultural production. The second example is the steam engine, which supported the transition from an economy where muscle power is the "prime mover" to an economy that is highly mechanized and energy-intensive. Although the concept of a "revolutionary technology" is not a standard one, I believe it is useful for making the point that not all GPTs are created equal. It is plausible that artificial intelligence will eventually emerge as another revolutionary technology by drastically reducing the role of human labor in economic production.

One important lesson from the study of previous revolutionary technologies is that they can facilitate large and long-lasting changes in economic and social trends. For instance, the rate of technological progress increased dramatically around both the Neolithic Revolution and the Industrial Revolution. A number of prominent economists have argued that AI-driven automation could lead to another increase of this sort. If AI does prove to be a revolutionary technology, then it could produce changes that are far more fundamental and far-reaching than anything policymakers have experienced.

General Purpose Technologies in History

General purpose technologies

The concept of a general purpose technology was first developed in the early 1990s by Bresnahan and Trajtenberg (1995). Their central idea was that some technologies simply matter much more than others. As they put it: "Whole eras of technological progress and growth appear to be driven by a few 'General Purpose Technologies' (GPTs)." The key features that distinguish these technologies from others are their unusually pervasive use, their tendency to spawn complementary innovations, and their large inherent potential for technical improvement.[1]

Some evidence for the notion that GPTs have outsized economic impacts comes from the history of total factor productivity (TFP) growth. TFP is a measure of how efficiently investments of labor and other resources can be transformed into goods and services that people wish to buy. It is also often used as a proxy for the rate of technological progress. Historical estimates suggest that, for at least the past century, TFP growth in countries at the economic frontier has come primarily through a small number of waves (fig. 4.1). A common view is that the waves are linked to the widespread adoption of new GPTs (Bresnahan, 2010).

Although Bresnahan and Trajtenberg focused on steam power, electricity, and computers, other authors have since proposed additional technologies as possible GPTs. Table 4.1 includes several of these suggested technologies, drawing from a list generated by Lipsey et al. (2005).

FIGURE 4.1 Waves in the American productivity growth rate, over the past hundred years, according to estimates in the Long-Term Productivity Database (Bergeaud et al., 2017). Note that a high-pass filter, with $\lambda = 500$, has been used to smooth out short-run fluctuations. The most recent wave shown here is normally attributed to the successful adoption of computers and the internet, two recent GPTs (Brynjolfsson & Saunders, 2009). Electrification and the internal combustion engine are often believed to have played outsized roles in the larger, mid-century waves (Bergeaud et al., 2017; Bakker et al., 2019).

Table 4.1 General purpose technologies from across human history, as identified by Richard Lipsey

Technology	First significant use
Domesticated plants	9000–8000 BC
Domesticated animals	8500–7500 BC
Smelting of ore	8000–7000 BC
Wheel	4000–3000 BC
Writing	3400–3200 BC
Bronze	2800 BC
Iron	1200 BC
Water wheel	Early Middle Ages
Three-masted sailing ship	15th Century
Printing	16th Century
Steam engine	18th Century
Railways	19th Century
Internal combustion engine	19th Century
Electricity	19th Century
Computer	20th Century
Internet	20th Century

The concept of a "general purpose technology" is used primarily by economists, with economic impacts tending to serve as the primary standard for inclusion in the category. However, as should be clear from this list, GPTs nearly always have spillover effects on military affairs and politics that are also highly deserving of attention.

The case of electricity

As a useful illustration of the effects a GPT can have across economic, political, and military domains, we can consider the case of electricity. Beginning roughly with the invention of the battery in 1800, or the electric motor in 1821, electricity began to find an increasing array of applications. Beyond enabling methods of long-distance communication like the telegraph, it served as an almost universally applicable method of transmitting energy from the engines or turbines that generated it to machines that could take advantage of it. Although the adoption of electricity proceeded gradually, in countries at the economic frontier, the first half of the 20th century was a period of dramatic "electrification." It became easy to transmit large quantities of energy into individual homes and businesses. Factories were also freed from needing to design their production processes around a single central engine.

We can see the effects of electrification, first, in early twentieth century productivity growth statistics for leading countries such as the United States. We can also see the effects in accounts of how daily life changed for the typical person, as new products like refrigerators, washing machines, lightbulbs, and telephones were introduced (Gordon, 2017). These changes significantly raised living standards, while also helping to reduce the domestic burdens placed on women and very plausibly accelerating their entry into the workforce in the latter half of the twentieth century (Coen-Pirani et al., 2010). Electronic communication technologies like the telegraph and radio also enabled stronger forms of mass political communication and centralized governance, perhaps most notably put to use in a number of mid-twentieth century totalitarian regimes. At the same time, electricity played a key role in enabling a "revolution in military affairs" due to the significant military applications of the radio, the spark-ignition engine, radar, and code-breaking computers (Krepinivich, 1994). The radio, in particular, was core to the *Blitzkrieg* tactics that Germany successfully implemented in the Second World War. In more recent times, electricity has of course enabled all modern information technology and its various impacts as well.

Modest beginnings

The long history of electricity also demonstrates a trajectory common to most GPTs: a GPT typically begins as something crude and only narrowly significant, then slowly achieves a larger impact through decades of technical improvements, costly investments, diffusion of knowledge, individual and institutional adaptation, and further invention.

The history of steam power is perhaps even more remarkable in this regard, given that nearly 200 years passed before it was used for much beyond pumping water out of mines (Von Tunzelmann et al., 1978; Smil, 2018). A similar trajectory has also played out more recently for the computer. In the 1940s, engineers at IBM could apparently see no use for more than a "half-dozen" computers nationwide (Cohen, Welch, Campbell & Campbell, 1999).[2] It was not until the 1990s, about a half-century later, that the introduction of computers had a

noticeable impact on economic productivity, changed most people's daily lives substantially, or were used pervasively in a major military operation (Brynjolfsson & Saunders, 2009).[3]

Revolutionary technologies

If we look back further into the past, then it becomes clear that some periods of technological change are more radical than the rest. Economic historians commonly cite the Neolithic Revolution and the Industrial Revolution as periods that involved unusually fundamental changes to the nature of economic production.[4] I believe it is useful to classify such periods as bringing *revolutionary change*. General purpose technologies that play prominent roles in supporting revolutionary change can then be classified as *revolutionary technologies*.

The Neolithic Revolution involved a transition from an economy largely based on nomadic hunting and gathering to an economy largely based on sedentary agricultural production. It occurred in Western Asia between approximately 10,000 BCE and 5000 BCE, with other regions following after delays of varying lengths. Because domesticated crops played an especially central role in this transition, it is natural to classify them as a revolutionary technology.

The Industrial Revolution involved a transition from an economy largely based on agricultural production to an economy largely based on industrial production and the provision of services. It occurred in Western Europe and the United States between approximately 1750 and 1850, with other regions again following, after delays of varying length. In one interpretation, the Industrial Revolution was an energy transition more than anything else (Landers, 2005; Smil, 2018). The region moved beyond primarily relying on organic sources of energy and material—such as grain, wood, and manure from grass-fed animals—and in doing so opened new productive possibilities. Energy-intensive machines have increasingly supplanted the muscles of humans and animals. Because the steam engine played a very important role in this transition, it can also be classified as a "revolutionary technology."[5]

Shifts in the human trajectory

Both the Neolithic Revolution and Industrial Revolution were accompanied by a number of dramatic and long-lasting changes in economic and social trends. In other words, beyond their qualitative impact on economic production, both these revolutions can be said to have shifted the human trajectory in significant ways.

Energy capture

Perhaps the most fundamental trend to emphasize is growth in "energy capture," a measure of the total amount of energy harnessed by humanity (Morris, 2013). This includes energy captured through food eaten by humans and domesticated animals, through fuel used to produce heat or power machines, and through the accumulation and alteration of physical materials. For obvious reasons, energy capture is associated with the capacity to support a large population, wage war, manufacture goods, travel, process information, and generally,

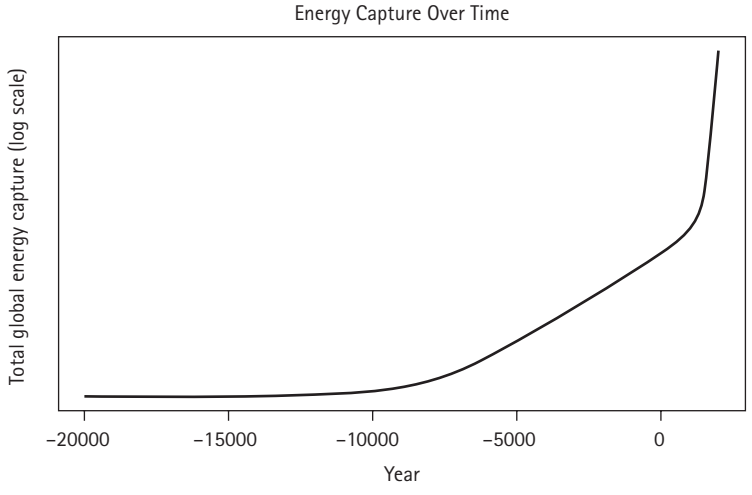

FIGURE 4.2 A stylized depiction of global energy capture over time. Although we lack reliable numerical estimates of the growth rate before the twentieth century, we can be fairly confident that it increased dramatically around both the Neolithic Revolution and the Industrial Revolution.

as the historian Ian Morris puts it, "get things done in the world" (Morris, 2010). Higher rates of energy capture growth have tended to reflect higher rates of technological progress and material change. Energy capture also lends itself more naturally to discussions of long-run trends than more sophisticated economic metrics. For instance, metrics such as gross domestic product and total factor productivity are not obviously applicable to hunter-gatherer societies.

Unsurprisingly, there are no reliable numerical estimates of global energy capture in pre-modern times. Nonetheless, we can be fairly confident that the rate of growth increased dramatically over the course of both economic revolutions (Smil, 2018). Sedentary farming societies, which manage dense concentrations of high-value plants and animals, extract far more energy per unit of land than mobile hunter-gatherer societies do. As a result, the spread and intensification of agriculture over several thousand years likely enabled an unprecedented rate of energy growth. The later transition away from organic sources of energy and toward fossil fuels, which began around the Industrial Revolution, then raised the rate of energy growth to even greater heights.[6]

Figure 4.2 presents a stylized depiction of global energy capture over time.[7] The events of the past century have all unfolded in the context of the nearly vertical section of the graph. When we looked at recent waves of technological change, this obscured the fact that even the lacunas between the waves contained unusually rapid change by pre-industrial standards. Ever since the Industrial Revolution, humanity's ability to "get things done in the world" has been growing at a rather exceptional pace.

Further shifts: The Neolithic Revolution

Most trend changes associated with the Neolithic Revolution are, unsurprisingly, fairly uncertain. Nonetheless, as groups of humans settled down and became farmers,

a very clear trend emerged toward larger, more complex, and more hierarchical social institutions (Morris, 2010). A typical pre-Neolithic group of hunter-gatherers probably included no more than a few dozen people. It also probably exhibited a relatively flat social hierarchy, with many important decisions being reached through rough consensus. The millennia that followed the Neolithic Revolution saw the emergence of increasingly large settlements, states, and ultimately empires. There was also a clear upward trend in the complexity of these institutions, for instance, in the volume of records maintained by states and in the sophistication of the military operations. An increasingly large portion of people came to live under the rule of powerful autocrats. Slavery also became increasingly prevalent.

Skeletal records and anthropological research suggests that, compared to hunter-gatherers, farmers tended to be more malnourished and sick but safer from interpersonal violence (Diamond, 1998; Wittwer-Backofen & Tomo, 2008; Eshed et al., 2010; Morris, 2014; Gat, 2017). In addition, it seems, farming societies tend to have stricter gender divisions than hunter-gatherer societies (Morris, 2015). One plausible partial explanation for this sharpening of gender roles is that the transition to farming increases the importance of upper body strength, for work outside the home, and men typically have more upper body strength than women. At the same time, women tend to become more tied to their homes, partly because sedentism allows for larger family sizes. Therefore, as agricultural practices diffused across the world, it is fairly likely that average living standards, levels of gender equality, and levels of violence all declined over thousands of years.

Further shifts: The Industrial Revolution

Perhaps the most notable outcome of the Industrial Revolution was the beginning of sustained growth in the average person's wealth (fig. 4.3). The current tendency for each generation to be noticeably wealthier than the generation that came before is historically anomalous. The transition to a much higher energy capture growth rate certainly played an important role in supporting the emergence and sustainability of this trend. Per-capita growth is only possible when total output grows quickly enough to outstrip the population growth rate.

The Industrial Revolution was also accompanied by a new trend toward more widespread democracy. Before the eighteenth century, democracy was an unusual form of government. Although some pre-modern states did have democratic elements, such as assemblies that constrained the actions of rulers, the ability of common people to participate was typically quite limited.[8] Furthermore, at the start of the eighteenth century, there was no clear sign of a global trend toward widespread democracy. Nonetheless, over the past 200 years, an obvious trend has emerged (fig. 4.4).

The Industrial Revolution was certainly not the sole cause of this new trend (Stasavage, 2020). A number of seemingly critical intellectual and political developments either predate the revolution or appear rather disconnected from it. The American Revolution, for instance, cannot be chalked up to industrialization. Nonetheless, many economists and historians do contend that the Industrial Revolution is an important part of the overall story. For instance, Acemoglu and Robinson (2006) suggest that agrarian economies are naturally less likely to democratize. So long as elites derive most of their income from rents on large tracts of land, they will have an especially strong reason to resist universal suffrage:

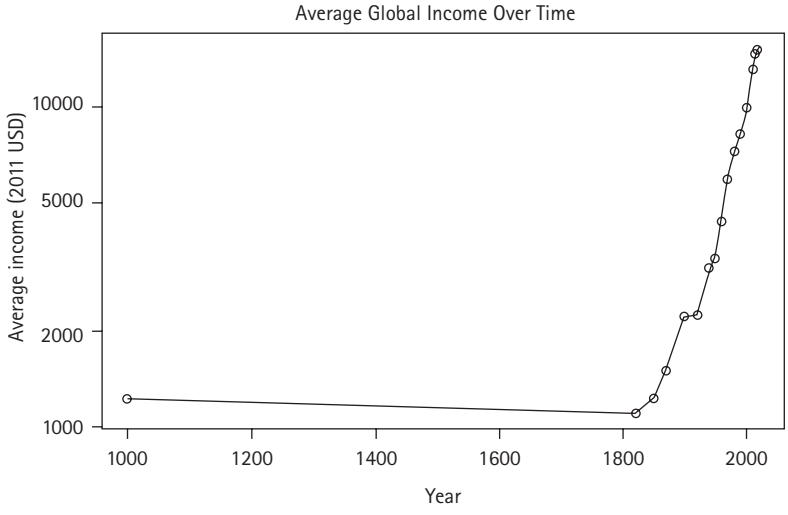

FIGURE 4.3 The average global income over time, according to estimates in the Maddison Project Database (Bolt & Van Zanden, 2020). The data point for 1000 CE is based solely on an estimate for China, which contained a large portion of the world's population at the time. Although the specific numbers given in the dataset are controversial, the qualitative story they suggest is not. There was no substantial, sustained, global income growth before the Industrial Revolution.

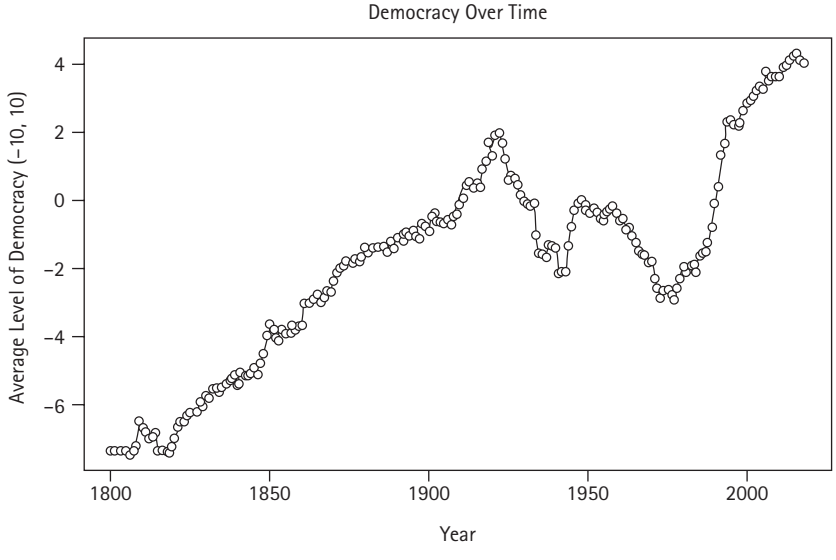

FIGURE 4.4 According to estimates in the PolityV dataset, the average level of democracy has been increasing for the past two hundred years (Marshall & Gurr, 2020).

ordinary people are likely to demand massive land redistribution. A population that is spread out across the countryside is also less able to coordinate and exert pressure on elites. Empirically, in modern times, there is also a clear statistical association between industrialization and democratization.

Other trends that appear to have emerged around the time of the Industrial Revolution, with at least a plausible causal link to it, include rising lifespans (Deaton, 2013), rising education (Lee & Lee, 2016), rising between-region inequality (Pomeranz, 2001; Deaton, 2013), rising gender egalitarianism (Morris, 2015), and declining rates of slavery.

Summary

In summary, throughout history, GPTs have tended to have broadly transformative impacts. However, these impacts have also tended to emerge quite gradually, typically taking many decades to become fully apparent. There are always frictions and barriers to impact that need to be overcome.

At least two GPTs, domesticated crops and the steam engine, stand out from the rest. These two *revolutionary technologies* both supported unusually fundamental changes to the nature of economic production. These economic changes in turn supported unusually significant and long-lasting shifts in the trajectory of humanity (Box 4.1).

When considering the future of artificial intelligence, it is useful to ask whether it will constitute a general purpose technology. We should also ask, though, whether it has the potential to become something even more transformative. If so, then popular analogies to recent GPTs such as electricity and the internal combustion engine may actually understate its ultimate significance.

Box 4.1

Some features of the world that we might use to judge revolutionary change are included in the following table. Upward-curved arrows indicate that a new or substantially faster upward trend likely began around the time of the Neolithic or Industrial Revolution. Downward-curved arrows indicate likely new or substantially faster downward trends. Blanks indicate that a trend did not change substantially or is highly ambiguous within a given period.

	Neolithic Revolution	Industrial Revolution
Total energy capture	⤴	⤴
Average income		⤴
Within-region income inequality	⤴	
Between-region income inequality		⤴
Average health	⤵	⤴
Gender equality	⤵	⤴
Information processing power		⤴
War-making capacity	⤴	⤴
Collective decision-making and political freedom	⤵	⤴
Degree of political centralization	⤴	⤴

Artificial Intelligence: Present and Future Impact

AI today

Today, artificial intelligence systems can perform only a very small portion of the tasks that humans are capable of. With some exceptions, existing systems also typically have narrow specialties: a typical system might just play a particular video game, recognize the faces of a particular group of people, or something of the sort. Although the term "artificial intelligence" sometimes evokes popular depictions of human-like machines, capable of the same sort of flexible reasoning that people engage in, nothing like this exists today.[9]

AI systems do already have a number of important applications. Nonetheless, at the time of writing, their significance still pales in comparison to the significance of previous general purpose technologies. This point holds regardless of whether we look at economic, military, or political impacts.

Economically, AI's most profitable present-day and near-term applications appear to be improved online marketing and sales systems, which make recommendations, select offers, and target advertisements in response to user data. Systems that help with supply chain optimization, for instance, by suggesting changes to order sizes and schedules, also appear to be highly valuable.[10] Other applications include more accurate fraud detection and more efficient data center cooling. While there are many other miscellaneous applications, such as the speech recognition software now installed on most smartphones, their total value to consumers still appears to be modest. Large investments have also been made into developing or exploring other more novel applications, particularly self-driving cars, but these are not yet in widespread use. Artificial intelligence has yet to have any clear impact on productivity growth, inequality, unemployment, or other macroeconomic trends (Brynjolfsson, Rock & Saunders, 2019).

In the military domain, its most valuable present-day applications may be associated with image recognition and the analysis of bulk-collected reconnaissance data (Pellerin, 2017). It is too soon to judge, though, just how much advantage will be gained through recent work. Autonomous military vehicles and much better systems for detecting cyber intrusions are also active areas of investigation, in a number of countries, but have yet to materialize or be widely deployed.

Politically, the most significant present-day applications of artificial intelligence may be related to law enforcement, online content recommendations, and target political advertising. Example applications of AI in the domain of law enforcement include predictive policing systems, which attempt to identify especially likely locations and perpetrators of crimes, and facial recognition systems, which can be used to identify and track individuals who appear in surveillance footage (Joh, 2017; Whittaker et al., 2018). A common hope is that these systems will help to reduce crime, while a common fear is that they will exacerbate or make it harder to reduce existing intergroup disparities in treatment by law enforcement. Some commentators worry that AI systems that recommend Facebook groups, YouTube videos, or other digital content based on users' behavior could have the effect of increasing political polarization or promoting false beliefs (Harris, 2020). A related concern is that

some targeted political advertisements could be substantially more effective or manipulative than typical political advertisements. Existing systems for generating convincing fake photographs and videos could also become politically significant, for instance if they are used to generate false depictions of political figures in compromising situations (Brundage et al., 2018). Again, though, most systems in this category are either not yet in widespread use or not yet effective enough to have had obvious society-level impacts on crime rates, voting patterns, intergroup disparities, levels of incarceration, and the like.[11]

Barriers to impact

Overall, the impact of AI has been limited in two ways. First, there are *technical bottlenecks* on what AI engineers can accomplish today. This means that existing techniques, sources of data, and quantities of available computing power are not sufficient to develop AI systems capable of performing many tasks of interest. The tasks that AI systems can perform today, such as targeted advertising and image recognition, possess a number of somewhat unusual traits that make them especially tractable.[12]

Second, there are *implementation challenges*. This means that even for potential applications that are not currently "out of bounds," the process of discovering, developing, and widely deploying these applications may be quite slow. Limiting factors include the scarcity of expertise, regulatory barriers, the unavoidable complexity of many engineering projects, the need to invent complementary technologies and services, and the need to attract large investments of capital. The relatively slow process of getting self-driving cars into widespread use has provided a clear illustration of several of these factors (Fagella, 2020).

AI's general purpose potential

Although its impact remains comparatively modest, artificial intelligence is a promising candidate for a new general purpose technology. As we have seen, it is already being applied in a wide range of domains. It is also inspiring enormous research and development efforts, which, by some estimates, might now account for substantially more than one percent of the world's total R&D spending.[13] Furthermore, if researchers can make enough progress on relevant technical bottlenecks, then the technology's potential for long-term improvement is enormous.

In fact, a growing number of economists have begun to identify AI as a likely GPT. This includes Manuel Trajtenberg, the founder of the GPT literature, and Erik Brynjolfsson, who is also a leading expert on the economic impact of information technologies (Trajtenberg, 2019; Brynjolfsson, 2019).

As discussed above, almost every GPT that has been developed so far began as something extremely crude with only a handful of practical applications. Artificial intelligence could then be in the early stages of an impact trajectory that several other technologies have followed before. In keeping with the two varieties of limitations mentioned above, there are roughly two reasons we might expect the impact of AI to grow over time. First, we might expect technical bottlenecks to decrease significantly over time, thereby "unlocking" many new applications.[14] Second, even without much of a change in capabilities, we can

reasonably expect many more applications to be developed and come into widespread use in the coming decades. The process of discovering what is already possible and implementing it could continue for a very long time.[15]

Economically, two especially significant applications that might be nearly within reach are self-driving cars and customer service systems. Given that over five million Americans are currently employed as vehicle operators or in call centers, Erik Brynjolfsson argues that automating a large portion of these roles over the next few decades would significantly boost national productivity and impact many workers' lives (Brynjolfsson, 2019). There has also been major recent progress in developing various kinds of *generative models*, such as systems that produce illustrations when given text prompts, that produce essays when given opening sentences, and even systems that produce lines of code when given descriptions of the desired code (Brown et al., 2020; Chen et al., 2021). At the time of writing, these systems are not yet ready for very widespread commercial use. However, progress continues to be rapid, and highly valuable applications might ultimately be very close at hand. A final promising application area to note is biomedical research. A recent major breakthrough in protein structure prediction suggests that artificial intelligence could find significant near-term applications related to drug design (Jumper et al., 2021).

Various authors have produced dramatic estimates of the portion of current jobs that could ultimately be automated, given present capabilities, often producing numbers in the double digits (Arntz et al., 2016; Chui et al., 2016; Frey & Osbourne, 2017; Winick, 2018). Some economists have also suggested, controversially, that such a wave of automation drawing on existing AI techniques could raise income inequality or unemployment.[16] However, there is still no consensus about the near-term economic impact of AI (Cukier, 2018).

In the military domain, present capabilities may be sufficient to develop greatly improved autonomous vehicles, systems for analyzing reconnaissance data, and systems to aid cyber offense and defense. Weaponized drone swarms, intended to overwhelm the defenses of large weapons platforms such as aircraft carriers, are one application that could have an especially large impact on the character of war (Scharre, 2014). There may also be valuable applications involving the acceleration or improvement of behind-the-scenes processes, such as vetting individuals for security clearances or determining vehicle maintenance schedules (DARPA, 2018). Ultimately, artificial intelligence could increase military effectiveness substantially. It might also shift important strategic parameters, such as the likelihood of accidental escalation, the relative ease of offense and defense, or the speed of military power transitions (Allen & Chan, 2017; Horowitz, Allen, Kania & Scharre, 2018; Scharre, 2018). Such shifts might then have an influence on the likelihood of war.

In the political domain, we might see continued improvement and adoption of systems for more effective law enforcement, political persuasion, and video forgery. Some authors have raised concerns that these applications could make open, fact-based political discourse more difficult and bolster authoritarian regimes (Brundage et al., 2018). More positive effects might include reduced crime or risk from terrorism due to law enforcement applications, or improved government decision-making due to better data-driven analysis. The possibility of increased inequality, unemployment, or international economic and military competition could also pose substantial political challenges (Dafoe, 2018; Horowitz et al., 2018).

Table 4.2 summarizes some existing and potential applications, including the several just discussed. Arguably, this list is still substantially less radical than the list of applications

Table 4.2 Several existing or emerging applications of artificial intelligence. All these applications might be feasible even without further progress on technical bottlenecks.

Economic	Military	Political
• Self-driving vehicles • Call center automation • Targeted advertising • Supply chain management • Protein structure prediction • Medical diagnosis • Electronic assistants • Text and speech translation • Text drafting • Code autocompletion • Illustration generation	• Drone swarms (reconnaissance or combat) • Image data analysis • Cyber intrusion detection • Software vulnerability discovery • Logistical optimization • Signal jamming • Facial recognition (counter-insurgency)	• Targeted political advertising • Image, video, and audio forgery • Facial recognition (surveillance) • Drone swarms (surveillance) • Predictive policing • Social media bots • Content recommender systems

that other GPTs, such as electricity, have already found over the past century. However, it is important to remember that GPTs tend to have a "long tail" of applications that were never anticipated in the early days of their development and use. There may be many important uses that have not been identified yet or that will only become feasible once some additional progress on technical bottlenecks is made.[17]

AI's revolutionary potential

Most researchers believe that AI systems will eventually be able to perform all the tasks that people are capable of. According to one survey, the median AI researcher even believes this milestone will probably be reached within the current century (Grace et al., 2018). Future generations might then live in a world of *near-complete automation*, meaning a world in which workers play little to no role in the production of goods and services. The simple logic of near-complete automation, as described by the economist Brynjolfsson & McAfee, 2014, is that potential employers have little reason to hire humans if it would always be cheaper and at least as effective to use a machine.[18,19]

Although near-complete automation would require tremendous progress on technical bottlenecks, to say nothing of implementation challenges, this scenario at least deserves serious consideration. The transition away from human labor would surely constitute another period of revolutionary challenge, alongside the Neolithic and Industrial Revolutions, and AI would surely constitute another revolutionary technology.[20]

No one knows what a world with near-complete automation would look like.[21] We cannot simply imagine our present world, only with some particular handful of tweaks and with everyone staying home from work. The differences would almost certainly be too vast for any detailed forecasting effort to be feasible. Nevertheless, there are still some positive predictions we can make. In particular, we can identify at least a few long-standing trends that are likely to break if we approach near-complete automation. Like domesticated crops and steam power, AI could facilitate major shifts in the human trajectory.

First, essentially by definition, there would be a sharp decline in the portion of people working for a living. If we include labor performed at home and in hunter-gatherer groups, then the expectation that most healthy people should work has been essentially a constant for all of human history. While labor participation rates do seem to be on the decline in many Western countries, the rate of this decline is fairly glacial compared to what would be implied by near-complete automation occurring within the century.

Second, growth trends might also be broken, just as they were by the development of agriculture and steam power. A handful of leading growth economists have recently begun to explore the implications of future automation for economic growth (Nordhaus, 2015; Aghion et al., 2018; Trammell & Korinek, 2020). They note that a number of commonly used models seem to suggest that approaching full automation should produce wildly accelerating growth rates. The validity of these models is still unclear. However, dramatically and meaningfully faster growth does at least seem to be a "live" possibility. This faster growth might be reflected in a broad range of metrics, ranging from GDP to global energy capture to various direct measures of technological progress.

Third, it seems plausible that inequality would rise rapidly for some time.[22] The possibility of a significant decline in labor force participation suggests that the vast majority of future income could take the form of returns on investments rather than wages. This means that the potentially enormous wealth generated through economic growth would accrue primarily to people who own significant capital, rather than to potential workers with little or nothing to invest (Hanson, 2016; Korinek & Stiglitz, 2018). Of course, redistributive policies could reverse this apparent consequence.

Fourth, we would have reason to expect a similarly sharp downward trend in the extent of humans' roles in warfare. Former U.S. Secretary of Defense James Mattis has even gone so far as to suggest that the automation of combat might be the first development to change the "fundamental nature of war" in all of human history (Mattis, 2018). Besides reducing the role of human combatants in an unprecedented manner, it is likely that the development of advanced AI systems would dramatically alter military weaponry, tactics, and strategy. It is also plausible that their development would grant militaries discontinuously greater capacities to cause destruction by reducing manpower constraints and accelerating the research and development of additional weapons.

Any of the above changes would surely have unprecedented political effects, many of which would represent breaks from existing trends. For example, we can attempt to imagine how the relationship between individuals and their governments would change if most people no longer worked for a living and if law enforcement could largely be automated. It seems plausible that under these circumstances the post-Industrial Revolution trend toward greater democratization would be reversed (Garfinkel, 2021).[23] Some authors have also suggested that sufficiently advanced AI systems or collections of AI systems could largely replace human governments in their roles providing services, enforcing rules, and making various decisions about how best to use resources. One way to frame this possibility is as a potential transfer of political power to trusted AI systems, which might ultimately be dramatically more capable than humans in many ways.

Finally, although extinction concerns are not a focus of this essay, it is certainly worth noting that some computer scientists and philosophers believe that the development of advanced AI systems could result in human extinction (Good, 1966; Yudkowsky, 2008;

Bostrom, 2014; Russell, 2019; Ord, 2020). If there is an intentional or unintended transfer of power to AI systems, then it may not be possible to take this power back. Furthermore, just as humans have threatened many other species by transforming and consuming resources they need to survive, it is conceivable that the behavior of these AI systems could be inconsistent with long-term human survival. Human extinction would, of course, imply a number of changes in current trends.

Conclusion

Compared to the most transformative technologies in history, artificial intelligence has not yet had a very large impact on the world. AI systems are still quite limited in their capabilities and many anticipated applications are still not ready for widespread use.

Nonetheless, there is a reasonable chance that artificial intelligence will emerge as a potent general purpose technology over the next several decades. It could have a range of economic, military, and political impacts that are at least comparable to those of computers or electricity in previous decades. Some of these impacts might manifest as notable changes in economic productivity, inequality, employment, health, safety, military effectiveness, or political freedom.

Furthermore, there is the more radical possibility that artificial intelligence will help to usher in a period of "revolutionary" change like what occurred during the Neolithic and Industrial Revolutions. This could imply dramatic shifts in many long-standing trends, unlike anything that has occurred in the past century and a half. Because the present pace of change is already so high, living through a transition of this sort could be in certain ways a more extreme experience than living at any previous time in history.

Notes

1. Of course, not all important technologies are general purpose technologies. Two examples to the contrary would be vaccines and nuclear weapons. These technologies had quite large impacts on certain metrics, pertaining to health and the potential destructiveness of war, but have had far fewer uses, are used in far fewer contexts, and have spawned far fewer innovations. One way to put this point more precisely is that these technologies have had only relatively *narrow* impacts. Removing vaccines and nuclear weapons from the world would seemingly require less individual and institutional adaptation than removing electricity would. On the other hand, removing electricity from the world would require dramatic adjustments in almost every domain.
2. Howard Aiken, the designer of IBM's Harvard Mark 1 computer, is said to have remarked: "[T]here was no thought in mind that computing machines should be used for anything other than out-and-out mathematics . . . . No one was more surprised than I when I found out that these machines were ultimately to be used for control in business . . . . Originally one thought that if there were a half dozen large computers in this country, hidden away in research laboratories, this would take care of all the requirements we had throughout the country" (Cohen, Welch, Campbell & Campbell, 1999).

3. One additional observation is that impacts on productivity and other formal measures of change often lag behind more visible indicators. In the case of both electricity and computers, the widespread adoption of the technologies by businesses occurred at least a decade before increases in productivity were observed. In the late 1980s, the ongoing lag led to discussions of a so-called "productivity paradox," sometimes summarized through the observation that "you can see the computer age everywhere but in the productivity statistics" (Brynjolfsson, 1993).
4. North and Thomas (1977) write: "Man's shift from being a hunter and gatherer to a producer of food has been regarded by common consent as one of the two major breakthroughs in his ascent from savagery to modern civilization [the other being the Industrial Revolution]." McCloskey (2014) writes: "[The Industrial Revolution] is certainly the most important event in the history of humanity since the domestication of animals and plants, perhaps the most important since the invention of language." A similar interpretation of the Neolithic and Industrial Revolutions as especially important transition periods in human history is also common among historians working within the framework of "Big History" (Christian, 2011). Morris (2015) and Hanson (2016) also find it natural to divide human communities throughout history into forager societies, farmer societies, and fossil fuel societies, with each revolution marking the emergence of a new kind of society.
5. Notably, the steam engine was not a major *cause* of the economic changes that occurred over the course of the Industrial Revolution. Economic historians now recognize that it was not widely used until the end of the period (Allen, 2009). It is instead more appropriate to think of the widespread use of steam power as a critical *outcome* of the Industrial Revolution, which cemented the revolution's long-run significance.
6. As an alternative historical narrative, Kremer (1993) suggests that it may be wrong to pick out the Neolithic Revolution and the Industrial Revolution as special periods in history when the rate of growth increased dramatically. He suggests that the growth rate may actually have followed a consistent (but noisy) upward trend between 1,000,000 BCE and the 1950 CE, rather than rising through two relatively distinct steps. Although the focus of Kremer's paper is population growth, rather than energy growth, population levels and energy capture are closely intertwined. See also Roodman (2020) for a discussion of this more continuous interpretation of historical growth.
7. This graph is loosely based on one classic attempt to estimate human population levels over time (McEvedy & Jones, 1978) and a more recent attempt to estimate frontier per-capita energy capture over time (Morris, 2013). Because these estimates are highly speculative, I have opted for a simple stylized graph over a graph that corresponds to a specific set of numerical estimates.
8. Stasavage (2020) distinguishes between "early democracy" and "modern democracy." Modern democracy involves recurring elections in which most members of a society are allowed to participate in the selection of its leaders. Early democracies fail to meet these criteria, but still nevertheless place constraints on leaders and allow for a significant degree of public participation. Stasavage notes that there have been many examples of early democracies, throughout history, but that early democracy tended to be confined to small states and that modern democracies essentially did not exist until recently.
9. In the context of this chapter, the term "artificial intelligence" refers to software systems that can perform complex cognitive tasks and that have been, at least in part, developed using machine learning techniques. Other authors, of course, use the term more or less inclusively.

10. One recent report, produced by the McKinsey Global Institute, predicts that online marketing and sales and supply chain management will together be responsible for the bulk of AI's near-term economic impact (Bughin et al., 2018).
11. Some commentators have suggested that artificial intelligence played an important role in the 2016 U.S. presidential election. In particular, a number of stories appeared shortly after the election claiming that British consulting firm Cambridge Analytica significantly skewed the results by applying voter targeting systems to large collections of Facebook data. However, there is little evidence that Cambridge Analytica or similar groups have had a large impact on voter behavior (Trump, 2018; Benkler et al., 2018).
12. Among other limitations, at the time of writing, it is still difficult to develop systems that can perform tasks involving very long sequences of decisions, tasks for which feedback is difficult to automate and for which good performance cannot be illustrated through many thousands of examples, and tasks that require the ability to adapt to highly dissimilar environments and unforeseen changes in circumstances (Brynjolfsson & Mcafee, 2017). Arguably, this set of limitations is relevant to the majority of tasks that people might want to automate. Consider, for example, the tasks "planning military operations," "writing bestselling mystery novels," or "cleaning houses."
13. Although investments into "artificial intelligence" are sometimes difficult to distinguish between more general investments into computer software and hardware, total investments into AI by leading technology companies are typically estimated to be in the tens of billions of dollars. For instance, a 2018 McKinsey report arrives at an estimate of between $20 and $30 billion per year (Bughin et al., 2018). The figure also very likely increased a great deal in the following years. A Congressional Research Service report estimates that total global R&D expenditures in 2018 totaled two trillion dollars (Congressional Research Service, 2021).
14. In fact, we have already seen substantial reductions in these bottlenecks. It is widely held, for example, that the "deep learning" techniques behind a large portion of recent applications only became fully practical within the past two decades. Key factors in removing previous limitations seem to have been increased computing power, larger data sets to learn from, and various algorithmic innovations (Goodfellow et al., 2016). Today, resources and research effort continue to increase rapidly and might continue "unlocking" new applications for quite some time. Robotics stands out as one particularly important area where capabilities are still fairly limited, but which continues to show interesting signs of progress (Andrychowicz et al., 2020).
15. The Chinese venture capitalist and former AI researcher Kai-Fu Lee, for example, is one notable proponent of the view that we may be entering a lengthy "age of implementation" with relatively little progress on fundamental capabilities (Lee, 2018).
16. Autor (2015) expresses a mainstream view on the possibility of technological unemployment. He notes that concerns about technological employment have arisen very many times in the past and have consistently been mistaken. Although Autor notes that there is no "fundamental economic law" guaranteeing that there will always be enough new jobs to replace old ones, he is skeptical that automation driven by emerging technologies will increase unemployment. Nevertheless, there is not yet an academic consensus on this question, and other economists, such as Acemoglu and Restrepo (2020), have expressed a higher level of concern about near-term unemployment effects. There is also some evidence that ongoing automation may be causing a decline in middle-skill jobs and an increase in the number of low-skill and high-skill jobs. This phenomenon is known as "job

polarization" and may be a cause of increasing economic inequality and slowing median wage growth (Acemoglu & Autor, 2011).

17. The economist Robert Gordon is perhaps the most prominent skeptic of the view that artificial intelligence will have an impact comparable to previous GPTs, at least in the economic domain (Gordon, 2017). He argues that the mid-century surge of productivity growth, which was in part driven by electrification and the adoption of the internal combustion engine, was an unrepeated and perhaps unrepeatable historical anomaly. In his view, the computer and the internet have been, overall, much less transformative and that artificial intelligence is likely to be even less transformative still.
18. It is important to emphasize that expecting artificial intelligence to eventually lead to widespread unemployment does not mean that we should expect it to increase unemployment in the short term. As discussed, the potential employment effects of present-day AI systems are a matter of significant controversy.
19. While there are some services that people might value specifically *because* they are performed by humans—such as professional sports and political representation—there would not be an obvious role for human labor outside of such exceptions.
20. The Open Philanthropy Project, a large philanthropic organization, has introduced the concept of "transformative AI" to discuss essentially this possibility (Karnofsky, 2016). They define "transformative AI" as "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution."
21. There are different visions of what kinds of AI systems might enable near-complete automation. Some authors, such as Bostrom (2014), suggest that highly general and agential AI systems might play crucial roles. Other authors, such as Drexler (2019), suggest that complex networks of narrow and tool-like systems, resembling the kinds of systems that are in use today, might be sufficient.
22. Perhaps counter-intuitively, the Industrial Revolution does not appear to have initiated any long-term trend toward greater inequality within countries (Bourguignon & Morrisson, 2002). Average within-country inequality also does not appear to be increasing today (Hasell, 2018). This means that rapidly rising inequality would in fact constitute a trend change.
23. A number of political scientists and economists have developed economic explanations of the circumstances under which democracy is likely to emerge and be sustained. Some of these explanations, such as those presented by Acemoglu and Robinson (2006), point at factors that would plausibly be diminished in a world with near-complete automation. For instance, they point to high inequality, fear of policies to redistribute factors of production, and vulnerability to pressure from workers as risk factors for dictatorship. Given that high levels of democracy are a somewhat anomalous feature of the post-Industrial Revolution era, it seems that, in general, we should not be too confident they will also be a feature of the next economic era.

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In D. Card, & O. Ashenfelter (Eds.), *Handbook of labor economics* (Vol. 4, pp. 1043–1171). Elsevier.

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society 13*(1), 25–35.

Acemoglu, D., & Robinson, J. A. (2006). *Economic origins of dictatorship and democracy*. Cambridge University Press.

Aghion, P., Jones, B. F., & Jones, C. I. (2018). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 237–282). University of Chicago Press.

Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., & Zaremba, W. (2020). Learning dexterous in-hand manipulation. *The International Journal of Robotics Research 39*(1), 3–20.

Arntz, M., Gregory, T., and Zierahn, U. (2016). *The risk of automation for jobs in OECD countries: A comparative analysis, OECD social, employment and migration*. Working Papers No. 189. Paris: OECD Publishing.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives 29*(3), 3–30.

Allen, G., & Chan, T. (2017). *Artificial intelligence and national security*. Cambridge, MA: Belfer Center for Science and International Affairs.

Allen, R. C. (2009). *The British industrial revolution in global perspective*. Cambridge University Press.

Bakker, G., Crafts, N., & Woltjer, P. (2019). The sources of growth in a technologically progressive economy: The United States, 1899–1941. *The Economic Journal 129*(622), 2267–2294.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bergeaud, A., Cette, G., & Lecat, R. (2017). Total factor productivity in advanced countries: A long-term perspective. *International Productivity Monitor 32*, 6.

Bolt, J., & Van Zanden, J. L. (2020). *Maddison style estimates of the evolution of the world economy. A new 2020 update*. Maddison Project.

Bostrom, Nick. (2014). *Superintelligence: Paths, dangers, and strategies*. Oxford University Press.

Bourguignon, F., & Morrisson, C. (2002). Inequality among world citizens: 1820–1992. *American Economic Review 92*(4), 727–744.

Bresnahan, T. (2010). General purpose technologies. *Handbook of the Economics of Innovation 2*, 761–791.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies "engines of growth"? *Journal of Econometrics 65*(1), 83–108.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Allen, G., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., . . . & Armodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv Preprint ArXiv:1802.07228.

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM 36*(12), 66–77.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Brynjolfsson, E., & Mcafee, A. (2017). Artificial intelligence, for real. *Harvard Business Review 1*, 1–31.

Brynjolfsson, E., & Saunders, A. (2009). *Wired for innovation: How information technology is reshaping the economy*. MIT Press.

Brynjolfsson, E., Rock, D., & Syverson, C. (2019). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In *The economics of artificial intelligence* (pp. 23–60). University of Chicago Press.

Bughin, J., Hazan, E., Lund, S., Dahlström, P., Wiesinger, A., & Subramaniam, A. (2018). *Skill shift: Automation and the future of the workforce* (pp. 3–84). McKinsey Global Institute.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., & Brockman, G. (2021). Evaluating large language models trained on code. ArXiv Preprint ArXiv:2107.03374.

Christian, D. (2011). *Maps of time: An introduction to big history* (Vol. 2). University of California Press.

Chui, M., Manyika, J., & Miremadi, M. (2016, July 8). Where machines could replace humans—and where they can't (yet). *McKinsey Quarterly*. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet.

Clifford, C. (2018, February 1). Google CEO: A.I. is more important than fire or electricity. *CNBC*. https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html.

Coen-Pirani, D., León, A., & Lugauer, S. (2010). The effect of household appliances on female labor force participation: Evidence from microdata. *Labour Economics 17*(3), 503–513.

Cohen, I. B., Welch, G. W., Campbell, R. V., & Campbell, R. V. (1999). *Makin' numbers: Howard Aiken and the computer*. MIT Press.

Congressional Research Service. (2021). *Global research and development expenditures: Fact sheet*. Congressional Research Service.

Cukier, Kenneth. (2018). The economic implications of artificial intelligence. In *Artificial intelligence and international affairs: Disruption anticipated* (pp. 29–42). Chatham House.

Dafoe, A. (2018). *AI governance: A research agenda. Governance of AI Program, Future of Humanity Institute*, University of Oxford, UK, 1442, 1443.

Deaton, A. (2013). *The great escape*. Princeton University Press.

Defense Advanced Research Projects Agency. (2018). *AI Next Campaign*. DARPA. https://www.darpa.mil/work-with-us/ai-next-campaign.

Diamond, J. M. (1998). *Guns, germs and steel: A short history of everybody for the last 13,000 years*. Random House.

Drexler, K. E. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence. *Future of Humanity Institute*. University of Oxford.

Eshed, V., Gopher, A., Pinhasi, R., & Hershkovitz, I. (2010). Paleopathology and the origin of agriculture in the Levant. *American Journal of Physical Anthropology 143*(1), 121–133.

Faggella, Daniel T. (2020). The self-driving car timeline—Predictions from the top 11 global automakers. *Emerj*. https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological Forecasting and Social Change 114*, 254–280.

Garfinkel, B. (2021). Is democracy a fad? *The Best That Can Happen*. https://benmgarfinkel.blog/2021/02/26/is-democracy-a-fad/.

Gat, A. (2017). *The causes of war and the spread of peace: But will war rebound?* Oxford University Press.

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers* (Vol. 6, pp. 31–88). Elsevier.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gordon, R. J. (2017). *The rise and fall of American growth*. Princeton University Press.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research 62*, 729–754.

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the Earth*. Oxford University Press.

Harris, T. (2020). *Americans at Risk: Manipulation and Deception in the Digital Age*. United States Congress (2020). Testimony of Tristan Harris.

Hasell, J. (2018, November). Is income inequality rising around the world. *World Economic Forum*. Available at: https://www.weforum.org/agenda/2018/11/is-income-inequality-rising-around-the-world.

Horowitz, M. C., Allen, G. C., Kania, E. B., & Scharre, P. (2018). *Strategic competition in an era of artificial intelligence*. Center for a New American Security.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 596*(7873), 583–589.

Joh, E. E. (2017). Artificial intelligence and policing: First questions. *Seattle UL Rev*. *41*, 1139.

Karnofsky, H. (2016). Some background on our views regarding advanced artificial intelligence. Open Philanthropy Project (Blog). https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advancedartificial-intelligence.

Karpathy, A. (2021, March 13). Software 2.0. *Medium*. https://karpathy.medium.com/software-2-0-a64152b37c35.

Korinek, A., & Stiglitz, J. E. (2018). Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence: An agenda* (pp. 349–390). University of Chicago Press.

Kremer, M. (1993). Population growth and technological change: One million BC to 1990. *The Quarterly Journal of Economics 108*(3), 681–716.

Krepinevich, A. F. (1994). Cavalry to computer: The pattern of military revolutions. *The National Interest 37*, 30–42.

Landers, J. (2005). *The field and the forge: Population, production, and power in the pre-industrial west*. Oxford University Press.

Lee, K. F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin.

Lee, J.-W., & Lee, H. (2016). Human capital in the long run. *Journal of Development Economics 122*, 147–169.

Lipsey, R. G., Carlaw, K. I., & Bekar, C. T. (2005). *Economic transformations: General purpose technologies and long-term economic growth*. Oxford University Press.

Marshall, M. G., & Gurr, T. R. (2020). *Polity 5: Political regime characteristics and transitions, 1800–2018*. Center for Systemic Peace.

Mattis, James. (2018, February 17). Press gaggle by Secretary Mattis en route to Washington, D.C. [Interview]. U.S. Department of Defense. https://www.defense.gov/News/Transcripts/Transcript/Article/1444921/press-gaggle-by-secretary-mattis-en-route-to-washington-dc/.

McCloskey, Deidre. (2004). Insight and notta lotta yada-yada: The Cambridge economic history of modern Britain. *Times Higher Education*. https://www.timeshighereducation.com/books/insight-and-notta-lotta-yada-yada/189766.article.

McEvedy, C., & Jones, R. (1978). *Atlas of world population history*. Puffin Books.

Morris, I. (2010). *Why the west rules-for now: The patterns of history and what they reveal about the future*. Profile Books.

Morris, I. (2013). *The measure of civilization*. Princeton University Press.

Morris, I. (2014). *War! What is it good for? Conflict and the progress of civilization from primates to robots*. Farrar, Straus and Giroux.

Morris, I. (2015). *Foragers, farmers, and fossil fuels*. Princeton University Press.

Ng, Andrew. (2017). *Artificial intelligence is the new electricity*. Future Forum, Stanford Graduate School of Business.

Nordhaus, W. D. (2015). *Are we approaching an economic singularity? Information technology and the future of economic growth (No. w21547)*. National Bureau of Economic Research.

North, D. C., & Thomas, R. P. (1977). The first economic revolution. *The Economic History Review 30*(2), 229–241.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Pellerin, Cheryl. (2017, July 21). Project Maven to deploy computer algorithms to war zone by year's end. U.S. Department of Defense. https://www.defense.gov/News/News-Stories/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/.

Pomeranz, K. (2001). *The great divergence*. Princeton University Press.

Roodman, D. (2020). *On the probability distribution of long-term changes in the growth rate of the global economy: An outside view*. Open Philanthropy.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Scharre, P. (2014). *Robotics on the battlefield part II*. Center for New American Security.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. WW Norton & Company.

Smil, V. (2018). *Energy and civilization: A history*. MIT Press.

Stasavage, D. (2020). *The decline and rise of democracy*. Princeton University Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Trajtenberg, M. (2019). Artificial intelligence as the next GPT: A political-economy perspective. In *The economics of artificial intelligence: An agenda* (pp. 175–186). University of Chicago Press.

Trammell, P., & Korinek, A. (2020). Economic growth under transformative AI. GPI Working Paper.

Trump, Kris-Stella. (2018, March 23). Four and a half reasons not to worry that Cambridge Analytica skewed the 2016 election. *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2018/03/23/four-and-a-half-reasons-not-to-worry-that-cambridge-analytica-skewed-the-2016-election/.

Von Tunzelmann, N., Von Tunzelmann, G., & Von, T. (1978). *Steam power and British industrialization to 1860*. Oxford University Press.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *The AI Now report 2018*. AI Now Institute.

Winick, E. (2018, January 25). Every study we could find on what automation will do to jobs, in one chart. *Technology Review*. https://www.technologyreview.com/2018/01/25/146020/every-study-we-could-find-on-what automation-will-do-to-jobs-in-one-chart/.

Wittwer-Backofen, U., & Tomo, N. (2008). From health to civilization stress? In search for traces of a health transition during the early Neolithic in Europe. In J. Bocquet-Appel & O. Bar-Yosef (Eds.), *The neolithic demographic transition and its consequences* (pp. 501–538). Dordrecht: Springer.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global Catastrophic Risks 1*(303), 184.

CHAPTER 5

AI GOVERNANCE MULTI-STAKEHOLDER CONVENING

K. GRETCHEN GREENE

Introduction

I spent a year at The Partnership on AI (PAI) in San Francisco, in 2019 and 2020, leading a multi-disciplinary, multi-stakeholder project on affective computing[1] and ethics, an experiment in AI governance. I had conversations with more than 200 engineers, scientists, lawyers, privacy and civil rights advocates, bioethicists, managers, executives, journalists, and government officials, mostly in the United States, United Kingdom, and European Union, about AI related to emotion and affect and its potential impacts on civil and human rights. I collaborated with the technical field's founder to increase the affective computing research community's focus on ethics and governance. I advised a prime minister's office and tech CEOs. I was interviewed on face and emotion recognition and U.S. artificial intelligence policy by international press, including the BBC and *Politico Europe*. I led multi-disciplinary convenings in London and Boston on affective computing and ethics and wrote *The Ethics of AI and Emotional Intelligence*.[2]

This collection of reflections from that year[3] draws lessons on convening, multi-disciplinary collaboration, and affective computing and AI ethics and governance. It offers a blueprint for creating the shared knowledge foundation non-technical participants need to apply their expertise. It presents question exploration as a tool for evaluating ethics risk and using "What is notice good for?" shows how a well-chosen question can serve as a catalyst for group or individual exploration of the issues, leading to insights and answers.

These questions and reflections on affective computing, AI, ethics, and governance are not just for multi-disciplinary AI ethics and governance conveners, but for all who are striving to find better ways to ask and answer questions together as a society, on AI governance or any other topic where no single actor has all the information needed to make the best decisions, the power to make every kind of useful intervention, nor the sole claim to be considered in the decision-making process.

The Many Roles of the Convener

Over the course of the year, I thought hard about what my role, my institution's role, and others' roles as multi-disciplinary AI and ethics conveners could and should be. I concluded that we should use our position at the center of an information and relationship network to identify participants' needs, priorities, and constraints to lower barriers to participation; increase the overall value created; understand the context, biases, and limitations of participants' contributions; and determine outputs, such as a checklist or lessons from a case study exercise, that could be usefully incorporated into existing industry AI production processes.

We must build a knowledge foundation for and with the participants, including shared terminology, a survey of current application areas, sensor types and kinds of inferences, and information about the accuracy and limitations of the technology, with sufficient detail to allow non-technical experts to apply their expertise and to understand possible sources of problems and points of intervention. We must build relationships, with participants and between them.

Finally, as conveners, we must develop tools and methods to create compelling discussions and lead the group to insights. I curated a list of questions for *The Ethics of AI and Emotional Intelligence*, captured from and inspired by conversations with participants. Questions like, "Who should make decisions about limits on AI?", "What is notice good for?", and "How might the widespread or cumulative use of the same or similar AI create societal changes and problems that any single use would not?" provide inspiration for analysis, an anchoring point for discussion in multi-disciplinary groups, and a glimpse into the conversations of that year.

I played a variety of roles as convener, from advisor to anonymizer, always seeking to reduce barriers and increase value.

The convener as advisor, mentor, and reputation builder

For some participants who were quite junior or who were trying to establish individual or organizational expertise and reputation in AI and ethics, I was well positioned to offer relevant knowledge and advice and to look for opportunities to help them increase their AI and ethics expertise and reputation within and beyond the convening and made a special effort to do so, including providing formal acknowledgment in written work and procuring budget and organizational approval to be able to serve as a formal supervisor to two junior participants looking for additional experience.

The convener as anonymizer and catalyst

One industry team used the affective computing and ethics project as a catalyst for internal thought and discussion about the topic, which they reported as quite useful. They also wanted to share their thoughts outside their team, especially for the benefit of smaller

companies with fewer resources, but only if they could do so anonymously. As convener, I introduced their ideas into the larger project discussion without revealing their origins.

What Participants Want and Need

Participants in the affective computing and ethics project wanted a wide variety of things,[4] none of which is specific to affective computing itself. They wanted practical tools, process improvements, and uniform standards; network, relationship, and reputation building; and information, ideas, and interesting discussions. They wanted a glimpse of the future. They wanted to be included, to ask for input from others, to share what they knew, to fulfil a sense of responsibility, and to contribute. They wanted to advance a specific agenda or position, to get their own terminology and practices adopted, to spread good practices, and to increase market share. They wanted to understand what good policy would look like and how to create a law that would support that policy. They wanted to keep their employers happy and build their own careers. They wanted anonymity. They wanted publicity. They wanted to improve outcomes from affective computing and AI applications for society.

Participants sometimes volunteered information about what they wanted. Sometimes I asked. Sometimes I inferred it from what they said and did.

Interests and concerns of technical and non-technical participants

In the broadest strokes, participants in the project could be put in one of two categories, technical or non-technical,[5] which seemed to predict a rough pattern of interests, concerns, and barriers to contribution, although there was variation within each group and some overlap between them.

Technical experts wanted a decision-making or process improvement tool, to broaden their discussions and networks, and to create guardrails for technology they had developed. One technical expert in industry proposed creating a practical tool that product teams could use to make decisions or improve their process, thinking perhaps a checklist or flowchart would be useful. Another highly valued the opportunity to meet and discuss affective computing and ethics with people outside their usual corporate and conference networks. A third voiced a strong sense of responsibility for having developed the technology, worried about its potential uses, and wanted to use their technical knowledge to help legislators craft good policy.

Non-technical participants questioned whether the technology worked, needed more information about the technology and its applications to apply their expertise, and were handicapped by public discourse that was plagued by language confusion and inconsistency. Influenced by communication problems in the broader public debate[6] and recent, well-publicized reports questioning the efficacy and scientific foundations of affective computing,[7] many non-technical participants' first question was whether the technology worked well and if it did not, thought there was nothing else to talk about. Non-technical

participants also often felt that they did not know enough about how the technology worked or where and how it was being used to be able to apply their expertise.

Technical and non-technical participants wanted specific examples to discuss and analyze, but they were also constrained, for different reasons, in the examples they could offer or consider publicly. Many non-technical participants could, at best, offer an example at the level of detail of a news article they might have read, and they wanted more detail than that. Technical participants, on the other hand, were often intimately acquainted with potential examples—products and research they were actively working on. But they and their companies would have had many good reasons, from trade secrets to public relations, not to want to make the details of those projects public or semi-public nor to want to encourage a public investigation of the projects' ethics shortcomings.[8] Participants from industry never proposed a detailed investigation of a specific product or research project they were working on although they did offer some projects they had been publicly associated with for inclusion in the high-level survey of applications being compiled for *The Ethics of AI and Emotional Intelligence*[9] and for examples of relevant AI and ethics improvements their companies had made.

Participants' motivations and experience affect the information they offer

Knowledge of participants' motivations is relevant to the convener's roles as recorder, synthesizer, moderator, analyst, and scholar. They need to consider how the motivations of different participants are likely to color the information provided and how they, the convener, should take that into account.

In the affective computing and ethics project, I saw overrepresentation of information and ideas that:

- Closely adhere to the organization's public position,
- Make the participant's organization look good,
- Promote an organization's current practices for broader adoption,
- Are not core to the participant's business (where trade secrets, confidentiality, public scrutiny, or disclosure of strategy are areas of concern), and
- Are influenced by the sources of information that the participant has easy access to (including their own work, newspaper articles, and the research papers and conference talks in their own field).

Motivations can be complex, and the same result can be reached in many ways. I saw that it might be in an industry participant's interest to promote one of their company's current practices related to AI and ethics for broader adoption for many reasons, including:

- The individual wants to contribute some idea in discussion, and this is a chance to do so.
- Someone asked how they did it at their company.
- This is the only way they have seen it done.

- They or their colleagues have thought about or tried other ways and they think this is the best way it can be done, or at least as good as any other.
- It gives them a competitive disadvantage on cost or other measures to do it unless others also do it.
- Change is hard, and it would be a lot of engineering work and/or politically difficult at their company to switch to a different method.
- It would be good for public relations to have others adopt their method or otherwise receive external validation that their methods are exceptional or at least sufficient.
- It would be easier to compare companies' performance on AI and ethics if they all used the same method, whatever it was (and they think their company would look good), so someone should propose a method for broad adoption.
- It would be easier to check the box that they had done the right thing on AI and ethics in their process if everyone agreed on what that was, so someone should propose a method for broad adoption.
- They believe that smaller companies have insufficient resources to spend time developing AI and ethics best practices and/or that there are great inefficiencies in every company trying to do it on their own and it is inexpensive for them to share their solution.

Building Relationships

At the beginning, a multi-disciplinary convener may be the only one with connections to the participants and it is easier to strengthen that hub and spoke network than to transform it into a lattice, where participants have strong relationships with each other, independent of the convener (see Figure 5.1).

But a hub and spoke network is vulnerable. When the project ends and the convener exits, the network will be completely disconnected, leaving all the participant nodes isolated. A lattice, with each node connected to its nearest neighbors, is far more resilient. If any single node or edge disappeared from a lattice, or if a few did, it would have little effect. A convener should strive to create a strong relationship lattice for the participants, with many connections for each, so that when the convener walks away, the participants can continue to develop multi-disciplinary perspectives on their affective computing and ethics work. In the future, in a good network, if an affective computing scientist has a question for a privacy lawyer, or a bioethicist wants a conversation with an engineer, they will have an acquaintance to call.

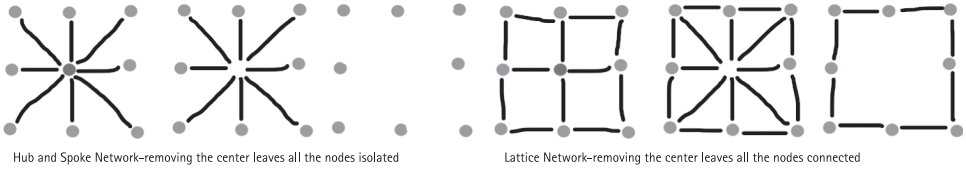
Hub and Spoke Network–removing the center leaves all the nodes isolated
Lattice Network–removing the center leaves all the nodes connected

FIGURE 5.1 AI Governance Multi-stakeholder Convening Hub and Spoke Network Graphic

Knowledge Foundation for Multi-disciplinary Discussions on AI and Ethics

The biggest challenge in creating fruitful, multi-stakeholder collaboration was to figure out what we could usefully do in a few hours or a day together or in conversations stretching across a year, what the barriers to progress were, and how we could overcome them.

The greatest barrier to progress was that there did not exist a resource to supply the knowledge foundation that would allow non-technical experts to apply their expertise and understand possible sources of problems and points of intervention. So, I wrote it.

The Ethics of AI and Emotional Intelligence addressed gaps in participants' knowledge, especially for non-technical participants, with the goal of creating the needed foundation for the multi-disciplinary, multi-stakeholder development of affective computing and ethics government policy and industry best practices. It provided a common language, addressing language use confusion in public discussion about affective computing; surveyed how the technology was being used; and provided questions inspired from the project's conversations and convenings to help direct discussion.

Its modified outline can be used for preparing the foundation for multi-disciplinary ethics and governance discussions in any area of AI.

Structure for an AI and ethics knowledge foundation

- Some basics of the technology
- Definitions (and communications/language issues)
- Survey of current applications (for the subset of AI of interest)
  - Sensors
  - Kinds of AI predictions
  - Industry/domain
  - Data subjects
  - How predictions are used/goals
- Information about the accuracy and limitations of the technology
- More detailed concrete examples for discussion
- Higher level AI and ethics questions that can be engaged with from a variety of backgrounds

Question Exploration as a Tool for Evaluating Ethics Risk

The convener has a major role to play in determining the direction of thought and discussion and helping lead the group to insights. Case studies are one good option, but there is

another worth exploring, a curated list of questions. I developed such a list for *The Ethics of AI and Emotional Intelligence*.[10] It serves as a partial record of the 2019–2020 affective computing and ethics project discussions; a distillation of what I found most salient; an open-ended starting point accessible to participants with diverse backgrounds; and a catalyst that can lead to a myriad of different lines of thought and insights, depending on the participants, the preceding conversation, and the day. The list and questions as I have used them in my work since, have evolved, and been adapted for more general AI use, but retain many of the ideas from the original list.

Thinking big

- What are the greatest benefits and opportunities we can imagine from this kind of AI?
- What are the greatest risks and harms we can imagine from this kind of AI?
- How is the use of this kind of AI impacting society?
- Does this kind of AI have applications that we should avoid completely?
- How might the widespread or cumulative use of the same or similar AI create societal changes and problems that any single use would not?
- Is this the right time for a broad discussion on ethics issues for this kind of AI?

How does this kind of AI fit into existing frameworks?

- How do concerns about this kind of AI fit within the larger debates about citizen, customer, user, and employee monitoring and surveillance?
- From what we know about other kinds of AI, what should we expect will go wrong for this kind?
- Does this kind of AI require different considerations or safeguards for development, deployment, procurement, or use, compared to other AI or other products or services?
- What types of laws are impacting this kind of AI and what safeguards have they created?
- What laws or types of laws should govern the data collection, inferences, and applications associated with this kind of AI?
- How should current events affect what we focus on when thinking about this kind of AI?

Human vs. machine

- What are the differences in what this kind of AI can detect or do compared to a human?
- How should human ability affect what uses or accuracy levels are acceptable for this kind of AI?
- How will embedding this kind of AI in robots impact human-human and human-robot interactions?

Accuracy and inclusivity

- How good is the technology and the science behind this kind of AI generally? How about in this use case specifically? Is it good enough to use in this application?
- How does the level of accuracy affect which ethics issues we should think about?
- How does accuracy for this kind of AI depend on the kinds of data, populations, predictions, and uses?
- How can we create and evaluate our technology and systems, so they work for everyone?
- When is it acceptable to build a product that only works for a subset of the population?
- How are the characteristics we are trying to measure, predict, affect, or simulate, with this kind of AI different in different cultures and how should that impact how we create or use the technology?
- How might internal team or external partner diversity improve subgroup accuracy or otherwise lower the ethics risk for this kind of AI?
- How could the inferences made by this kind of AI negatively impact or reveal a status historically protected from discrimination under the law, such as race, sexual orientation national origin, religion, gender, disability, or age?

Privacy and other rights

- How does this kind of AI impact an individual's ability to decide whether and when to reveal certain information?
- If this kind of AI reveals information the data subject wants to conceal, could it violate civil rights to freedom against unreasonable search and seizure or self-incrimination, human rights to freedom of thought and opinion, or other rights to limit access to sensitive health information or other information that should have special protection?
- Does this application impact an important right, like access to work, education, or housing?
- Can the data needed to use this kind of AI, or created by it, be used in ways that threaten other rights like freedom of speech, religion, or assembly?
- How can this kind of AI's ability to measure, classify, or influence create risks to privacy even if the software never has access to identifying or identifiable information?
- What data should be held only on a user's device?
- Who has access to the inferences this kind of AI makes about the data subject?

Best interest vs. autonomy

- Who should make decisions about limits on AI?
- When is "serving the best interests of the data subject" the right standard when designing and deploying this kind of AI?
- To what extent should those affected by this kind of AI be in control over when and how it is used, and who will ensure that they are?
- How can we give data subjects agency over inferences about themselves?
- When could telling data subjects inferences about themselves be harmful?

Communications and transparency

- How can we build a common knowledge base to facilitate discussions about this kind of AI across industry, academia, governments, news media, civil liberties organizations, and the general public?
- What information would someone need to challenge decisions based on this kind of AI?
- How does the language we use for this kind of AI affect beliefs about appropriate use and efficacy?
- What problems arise if the entity that deploys this kind of AI or the data subject does not understand what it actually measures?
- When will the developers themselves know what a deep learning model is implicitly recognizing or responding to?
- What should communication standards or best practices be for this kind of AI?
- Would notice create a meaningful opt-out for data subjects in the proposed uses of this kind of AI? What is notice good for?

Three Catalysts for Conversation

I have found the questions from *The Ethics of AI and Emotional Intelligence* to be useful prompts in my own writing and thinking, as a scholar and opinion writer, and as a kind of checklist to use in my work as a big technology company product consultant, evaluating AI and ethics issues for new products. I will use some of them in the next AI and ethics convening or training that I run and will continue to add questions and refine the list as I see which questions resonate with different groups and generate the best discussions.

As models of the kind of thinking that this kind of question might inspire in group discussion or for an individual writing exercise, here are three questions and where thinking about them led me one day, each blending smoothly into the next. Of course, any one of these questions might lead in a very different direction, depending on the group, the individual, the day, and how much time there is to think.

- Who should make decisions about limits on AI?
- What is notice good for?
- How might the widespread or cumulative use of the same or similar AI create societal changes and problems that any single use would not?

Who should make decisions about limits on AI?

There are a number of stakeholders that at least some of us might think should get to decide or get a say on data collection and use and AI, on what the limits should be, and how they should be enforced. We might say it should be the data subject or their legal guardian, the government, the company that built the technology, or the one that deployed it. But before asking who it is, we should ask why we think so. Asking what the source of their authority

or right would be or what other factors we should consider can help us decide whom to put on the list and how to prioritize their claims.

Sources of authority or right for making decisions about limits on AI

- Who has interests or rights that should be protected?
  - Who is affected and how much?
  - Who is the data or inference about?
  - Who is the creator?
  - Who has a property interest?
- Who has a duty to protect and make decisions for one or multiple other parties?
  - For one or a few (parents, guardians, caretakers)
  - For many (governments, schools)
- Who is well positioned to decide?
  - Who has the relevant information?
  - Who can get the relevant information held by different parties?
  - Who could understand and use all the information if they had it?
- Who is well positioned to enact controls?[11]
  - Who has direct control over the points of intervention?
  - Who has indirect control, such as regulatory or market power, over the parties with direct control?
- Who can do it efficiently?
- Who can be trusted?
- Who can be held accountable?

Working through the list of factors, it might be the data subject because it affects them, because it is about them, or because they have a property right in the computer or phone used to collect the data or the location where collection occurred. It might be the government because it has a duty to represent the interests of all or because it can be more efficient than many individuals acting independently. It might be the companies collecting data and creating or using AI because they are the creators, because they understand it best, because they control many of the points of intervention, or because they have a property right in the technology, the collected data, or the real or personal property where the data was collected or stored. Or it might be someone else.

We probably aren't trying to identify a single actor, but rather to resolve the tensions between the factors and combine information and power held in different hands. The parties in the best position to act may be the least trusted. It is often inefficient to make everyone make a choice, even if we believe they should be empowered to make one. No single actor has all the information needed to make the best decisions, the sole claim to an interest in making them, or the ability to enact every useful type of control.

Even if in the end, a single actor makes a final decision, three key questions will need to be addressed. How can the necessary information be brought together, how should different parties' interests be balanced, and what kinds of control are available that could be used to create the desired limits?

One of the most common controls in data protection law, one presumably meant to put power in the hands of the data subject, is notice. We will ask what notice is good for, who the intended audience is, and by what mechanism it might achieve what goals. We will examine the case of the ubiquitous website privacy policy.

What is notice good for?

Asking what purpose notice to all users serves will lead us to realize that the audience may not be who we thought it was, and that counterintuitively, notice read by almost no one may still be serving useful goals, efficiently reaching the few users that want to read it, when we don't know who they are, or if read by no users, still enabling advocates to challenge outcomes after the fact or push for policy change. Considering notice's apparent function as an effective user opt-out option for those who do read it, by allowing the user to choose substitute goods or services with preferable data policies or to be careful about what they project or reveal, we realize the high cost or impossibility of opting out if there really is no good substitute. Even if the seller does not hold a near monopoly, the widespread use of similar AI technologies and similar data policies will have the same effect, eviscerating notice's opt-out function.

Website privacy notices are read only by a few

Website privacy policies are the everyday example of notice that surely suggests to most users that notice is literally a waste of time, the time it takes to click through the clickwrap we aren't reading. According to a 2019 Pew Research survey, 25 percent of U.S. adults are asked almost daily to agree to some company's privacy terms.[12] Only nine percent say they always at least glance at them before agreeing (see Table 5.1).[13]

Without delving into legislative histories, legal requirements for easily understandable, detailed website privacy policies with prominent links or that force a user to open, scroll, and click to establish consent before allowing access to a page, would seem to exist to ensure that all users read the privacy policies and thereby understand how information about them is being collected, used, and shared. If that is the goal, privacy policies are failing miserably.

Notice requirements can achieve useful goals without being read by any of the data subjects

But there are goals they could be meeting, even though most users never or rarely read them. Perhaps users, or at least the full set of users, are not the intended audience. We consider what makes notice useful to users and other data subjects, especially in the hands of others.

Notice can be a tool for third parties acting at least partially in data subjects' interests, influencing future data collection policies and practices by providing a starting point for privacy advocates, proactive and influential users, journalists, academics, and government officials trying to understand and influence how data is being used.

Table 5.1 "Website privacy policies"

Website privacy policies
Potential audiences, goals, and mechanisms for achieving them

Potential Audience	Effect on User/Data Subject		Effect on Other	
	Notice/ information goals/effects	Action/choice goals/effects	Notice/Information goals/effects	Action/choice goals/effects
All users/data subjects	Directly inform user/data subject: What information about them is being collected? How is it being used? Who is using it? Make it easy for users to understand	Directly inform user/data subject about any rights or choices they may have (e.g., deletion or do not sell rights) Directly provide users with a method/information about a method for exercising a right (e.g., "Send an email to this address") Make it easy for users to exercise their rights or opt out by not using service or product	Company could collect information about how many users open privacy policy or make other choices	With information about user privacy actions and choices, company might change their privacy practices
Only some users/ data subjects	Same as for all users, but only for a subset (e.g., Only those who want to know)	Same as for all users, but only for a subset	Company could collect information about how many users open privacy policy or make other choices	With information about user privacy actions and choices, company might change their privacy practices
Third parties Privacy advocates, researchers, journalists, legislators, and others.	User may get informed by third parties directly or receive additional information as a result of third-party actions (e.g., lobbying for legislation with additional notification requirements)	User's possible actions may be affected by third parties' actions (e.g., lobbying for legislation with restrictions on data use could result in some products or features being unavailable to user)	Inform privacy advocates, researchers, journalists, legislators, and other third parties about what information about users is being collected, how it is being used, and who is using it	Third parties can inform or influence: users/data subjects, academic research, public opinion, and government policy

It can also be used by lawyers and other advocates to challenge decisions that used the data, potentially changing how the data collection and use impacts the data subject. The kinds of sensors and data that was collected, the type of analysis, and how it was used, typical elements of notice requirements, all offer potential avenues to argue that the decision process was faulty or unfair.

Notice to all may be the most efficient way to get notice to the few

Broad public notice makes the information easily available to the other actors already discussed, who can more efficiently take certain kinds of action than most data subjects can. It is also a very efficient way to get notice to an unknown subset of the data subjects. Maybe some of them need or want notice but only some of them do and it is hard to identify who does in advance. We can tell all of them where the detailed information is and let them self-select and read it if they want to.

Notice is valuable to the extent that it facilitates action

Transparency about how data is being collected and used is often talked about as if transparency had value on its own and data protection laws generally include a notice requirement. But even when notice is received and understood by data subjects, it is not necessarily useful. There is emotional value in knowing good news, even when disconnected from possible action. There is dignitary value in knowing that information was not withheld. But the practical value of information to data subjects depends in large part on how they can make use of it to act, and on the power and desire of third parties to use the information to take action on their behalf.

Notice may provide an effective opt out, allowing data subjects to avoid or confound data collection or analysis

Notice is valuable if it informs a choice the data subject is going to make, in time for them to incorporate the information and change their course of action. The most obvious choice available to a data subject, reading a privacy notice with terms they do not like, is to walk away.

Where there are suitable alternate travel routes or substitute goods or services, informed data subjects might use notice to opt out entirely by avoiding a physical or virtual area where sensors are collecting data for analysis or by choosing a competitor's product that comes with different privacy terms.

Or, instead of complete avoidance, a data subject might alter their behavior while in the area, to avoid or reduce the effectiveness of either the initial data collection or the analysis. When someone does not want to share certain information with people nearby, they might speak quietly or not at all, display facial expressions that are misaligned with their emotions, shield their faces or eyes from view, or try to show less interest than they feel in objects they may want to buy. These strategies and others, like traveling without a smart phone, turning certain computer settings on or off, or using adversarial attacks against computer vision programs, might be used against AI and sensor-based systems to opt out of successful data collection.

The cost of opting out may be very high

However, it is often not as easy as it sounds. Even with full information, there are limitations on the ability of data subjects to avoid data collection, send false signals, or find comparable substitute goods or services. There is also always a cost to evaluate options, switch to some alternative, or change individual behavior, and the costs can be prohibitively high. The result is that even though notice plus continued use of a place, goods, or services sometimes comes quite close in effect to a consent regime, providing a meaningful opt out option for data subjects, more often the regimes look quite different.

To understand how costly opting out would be for a data subject, it is important to understand the environment the data subject is operating in, and what their alternatives are.

How might the widespread or cumulative use of the same or similar AI create societal changes and problems that any single use would not?

There is nowhere to opt out to

One approach to data protection is to say to the data subject, "If you don't like it, go somewhere else," where somewhere else means somewhere with similar goods or services but different data policies. But when a company has a near monopoly or a particular policy or practice is widespread, the cost of going somewhere with a different policy or practice increases dramatically and it goes up more the more rare or essential the goods or services are.

Consider a city block with many similar clothing stores. If a single store posts a sign at the entrance saying that they use face recognition for fraud prevention, a potential customer wanting to avoid face recognition can turn left or right and find similar clothing to buy. But if every clothing store on the block uses face recognition, or if the city has only one clothing store or only one with the kind of clothing the customer wants or needs, the customer is left without a good alternative.

If the type of goods or services that are unobtainable without encountering face recognition (or any other kind of data collection and analysis) is a luxury goods category like cigarettes or yachts, society may consider it acceptable to allow all sellers to bundle data collection with the sale of goods, forcing the buyer to choose between data protection and acquisition of the goods. However, for essential goods, like groceries or clothing, consumers cannot completely opt out of the category, so allowing all sellers to bundle data collection with the sale of goods effectively means that there is no limit on the data collection. According to a Pew Research study, six in 10 Americans say they do not think it is possible to go through daily life without having their data collected.[14]

Homogeneity makes it uniformly bad for anyone it is bad for

Another impact from the cumulative use of the same or similar technology is that the disadvantages or unfavorable results are especially concentrated. Homogeneity is uniformly bad for anyone it is bad for.

An error in a single system, directly affects only those that use that system, and probably only some of those. With different systems using different code bases and training sets, we would expect different errors in each system.[15] But if a single system's use is widespread, so are its errors. Instead of different errors in different systems affecting different groups, the same error hits the same group again and again and again. Just as the use of FICO scores expanded from lending into home rental and job applications, decision making processes could become uniform in a way that would make it much more costly to get a negative decision, with the same result guaranteed not only everywhere in a single sector, but across domains.

This is problematic not just for errors. The magnification applies to all negative results.

There is arguably a benefit to society, or at least to individuals, in some variation in judgment.[16] Individuals do not need to appeal to every landlord, employer, and prospective friend or partner. If we get turned down, we have a chance to try somewhere else. If we make a mistake, or a decision maker does, we have a chance to go somewhere else and get a fresh start. While there are significant limitations in our ability to find people that will judge us differently or to get a fresh start in society, if the same program, algorithm, or trained machine learning model is deployed broadly, even universally, it will make it that much harder.

Analysis of a single product is often not enough to understand the effects of scale and uniformity.

These are two examples of a class of societal changes and problems created by the widespread or cumulative use of the same or similar AI. The AI and ethics analysis for a single product or deployment is usually about just that one product or deployment, and that seems like the natural scope for a product team or procurement officer to be considering, but it is not enough. If every product team considers their product in isolation, there are societal effects that they will not see, that their products all together cause.

Conclusion

I have tried to pass on lessons learned while designing and leading a multi-disciplinary, multi-stakeholder project on affective computing ethics and governance in 2019–2020, to benefit those who will follow a similar path.

Armed with an understanding of technical and non-technical participants' interests, the importance of strong cross-participant relationship building, and the many roles of conveners; a blueprint for the creation of a knowledge foundation that will allow non-technical participants to apply their expertise; and a set of questions to provoke discussions and insights; a new convener will be much better prepared to lead a multi-disciplinary project on AI ethics and governance. An experienced convener will learn something new.

For those who are looking for catalysts for AI and ethics discussion for industry, university, news media, or policy teams, or for your own writing and thinking, to take you a step beyond or to the side of what you have been thinking about, there are 42 questions to choose from.

If nothing else, I hope you enjoy the memory of a short exposition on the possible goals of notice, the next time you fail to read yet another website privacy policy.

Notes

1. While affective computing is discussed here as if it is a subset of AI, the definition from the founder of the modern field is broader (not all computing is AI). "Affective computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena" (Picard, 1997).
2. Greene, 2020.
3. The author's reflections, opinions, and ideas have also been influenced by other work that the author has done, during that year and before and after, including as Fellow at Harvard Kennedy School's Belfer Center, Senior Advisor at The Hastings Center, Assembly Fellow and Research Affiliate at Harvard Berkman Klein Center and MIT Media Lab, Founder and CEO at Greene Strategy and Analytics, Mathematician at Lawrence Livermore National Lab, Autonomous Vehicle Engineer at NextDroid, Computer Vision Engineer at focal, Associate at Ropes and Gray, and Partner at Baker Thomas Oakley Greene. They do not reflect the views of any institution that the author is now or has been affiliated with.
4. This list is largely based on the author's observations and inferences, but some are based on explicit statements about values from participants, made in conversations with the author in 2019–2020.
5. Technical is defined here to mean someone who writes or has written AI or other software code including software, computer vision, and machine learning engineers, data scientists, and researchers and scientists in industry and academia who develop and use AI.
6. There was significant inconsistency in language use causing frequent confusion about whether, for example, an author was talking about internal emotional states or perceived emotional expressions. See the discussion of language problems in affective computing discourse in Greene (2020, p. 5).
7. Barrett, L. F., et al., 2019, pp. 165–166.
8. Participants did not state why they could not use their own current projects as examples or that they could not, but these are two of the factors that I believe would have kept them from doing so, if they had wanted to.
9. While broad, the survey of affective computing applications in *The Ethics of AI and Emotional Intelligence* is somewhat skewed away from a certain kind of application, those that were harder to get or use information about, for a variety of reasons.
10. Greene, 2020, pp. 18–21.
11. Being well positioned to enforce rules does not necessarily imply that a party should have any right to help decide what the rules are, but it might, at least in the details. If, for example, two different engineering methods would support a policy equally well, it seems right to give the engineering company that knows the most about them and will pay to implement them, some power to help determine the choice between the two.
12. Survey conducted June 3–17, 2019 (Auxier, B., et al., 2019, p. 5).
13. Nine percent say they always read privacy policies before agreeing to them and among the 60 percent of U.S. adults who say they ever read privacy policies before agreeing to them, 22 percent say they typically read it all the way through, 35 percent say they read it part of the way through, and 43 percent say they glance over it (Auxier, B., et al., 2019, p. 38).
14. "A majority of adults (62%) do not think it is possible to go through daily life without having their data collected by companies, and 63% think the same about government data collection" (Auxier, B., et al., 2019, p. 30).

15. Although with different code and training data, we would still expect related errors if they result from something common to both programs, such as a similar level of underrepresentation of dark skinned faces in both training sets, caused by the same societal influences.
16. There are clearly disadvantages to variations in judgment where the variation is across society and certain groups or individuals disproportionately get the bad results, or if the variation in how a single individual will be judged by those they meet each day has such extremes that they are likely to be hurt or killed in a bad reaction. But if the consistent reaction someone gets is a just barely rejection, increasing the variance would put almost half of the reactions to them as just barely acceptances. If the consistent reaction is a middle range acceptance, increasing the variance drops some middles down to low acceptances but bumps others up to very strong.

References

Auxier, B., et. al. (2019). *Americans and privacy: Concerned, confused and feeling lack of control over their personal information*. Pew Research Center.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Corrigendum: Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest 20* (3), 165–166.

Greene, K. G. (2020). *The ethics of AI and emotional intelligence*. The Partnership on AI. Retrieved from https://perma.cc/6WJ2-BG94.

Picard, R. W. (1997). *Affective computing*. The MIT Press.

SECTION II

VALUE FOUNDATIONS OF AI GOVERNANCE

JOHANNES HIMMELREICH

CHAPTER 6

FAIRNESS

KATE VREDENBURGH

Introduction

For many, algorithmic decision-making mediates their access to credit, education, and employment; how they access information; the healthcare or state benefits they receive; and their coercive treatment by the state. If properly governed, AI could contribute to more evidence-based, consistent decision-making in the institutions that seriously and unavoidably shape people's lives. But, as AI is seriously re-shaping our individual and collective lives, academic research, journalistic investigation, and user reports have raised serious, repeated concerns of *algorithmic bias*, or the deviation of an algorithm's performance from what is required by some standard.[1] "Algorithmic bias" can mean statistical bias; that is, an algorithm's predictions deviate from the true state of the world, where this is detected using previously observed testing data. Here, I will be using "algorithmic bias" in a moral sense. Algorithmic bias is the encoding of wrongfully discriminatory social patterns into an algorithm.[2] This encoding usually occurs through the statistical regularities that the algorithm uses for its predictive or classificatory task.[3] Such bias has been demonstrated in facial recognition systems,[4] risk assessment tools in criminal justice,[5] tools to prioritize scarce healthcare resources and that mediate access to jobs,[6] education,[7] and credit[8]—to name a few recent and startling examples.

The problem of algorithmic bias is a difficult technical and philosophical problem. Bias is required to make any prediction, yet it also plagues any prediction problem—in making an inference from the past to the future, the decision-maker needs to make simplifying assumptions about what the world is like, but those simplifying assumptions can introduce inaccuracy into predictions.[9] Bias is just as serious a concern when a machine is making the inferences. Models developed through machine learning, for example, are a powerful decision aids because they use past data to discover predictively powerful correlations between available data and the target variable of interest. But, because of a history of discriminatory institutions and social practices, social identity properties are often powerful predictors of the outcome of interest. We thus need to be concerned about how bias can enter into the AI systems that are used as decision aids, or that automatically execute decisions on the basis of their predictions. In the next section, I expand on the problem of algorithmic bias, focusing on different places that bias can enter into the data science pipeline.

Despite research, political scrutiny, and activism over the last decade, algorithmic bias remains trenchant. One likely cause of this trenchancy is inaction that favors elite interests. Elites and members of privileged groups have an interest in maintaining power and privilege, and structural and individual discrimination can be a means to that end.[10] Discrimination in education[11] and in the private sector[12] shapes who develops and researches technology, leaving places with the most money and prestige disproportionately white, male, and WEIRD (Western, Educated, Industrialized, Rich, Democratic).

However, there is another problem: people do not actually agree on what the problem of algorithmic bias is. And, until there is mutual understanding between those who disagree about the problem, we will continue to lack regulatory or industry benchmarks to determine when AI-based decision-making is wrongfully biased.

Some take the problem to be a matter of *justice*, or the moral norms that govern our basic societal institutions, including norms of distribution and norms of respect. Others take the problem to be a matter of *fairness*, or whether alike individuals are treated equally. Fairness is one of the moral values that ought to shape our basic societal institution, but it is not the only value of justice. This disagreement has shaped debates over technical, regulatory, and governance interventions to reduce algorithmic bias.

I will lay out the problem of algorithmic bias, and different places where bias can enter in the development of an AI system. Then, I argue that the problem of algorithmic bias is sometimes conceived as a problem of justice, and sometimes as a problem of fairness. I illustrate this claim with a discussion of fairness metrics. Next, I argue for the priority of justice over fairness. I conclude by examining five governance strategies for algorithmic fairness and justice.

Algorithmic Bias

AI Now's 2019 *Report* lists inbuilt algorithmic bias as one of the emerging and urgent concerns raised by the deployment of AI systems into sensitive social domains.[13] AI systems raise such serious concerns about discrimination because, as we previously discussed, AI has serious impacts on people's lives, in domains such as healthcare, criminal justice, finance, and education. It is important to consider potential discrimination throughout the process of building AI.[14]

Why is algorithmic bias such a pressing and pervasive social and political problem? One important explanation is the historical and current global severity and prevalence of identity-based oppression. Racism, sexism, classism, ableism, and other identity-based oppressions can be thought of as technologies that govern the distribution of advantage, and the imposition of harms and disadvantage, through institutional and social practices.[15] And they are particularly powerful technologies, shaping where we live, where we work, with whom we associate, and so on.[16] Because AI systems are created by human beings and embedded in our institutions and everyday social practices, we should also expect that identity-based oppression also shapes these technologies.

This expectation has been borne out, as the previous examples show. But we also have special reasons to be concerned about AI systems, beyond the pervasiveness of discrimination. AI systems, especially those developed through machine learning, are an especially

powerful mirror of the past, as has been the focus of much recent scholarship.[17] AI systems, especially those generated through machine learning, are powerful decision-making tools for solving classification problems because they learn predictively useful patterns relevant to variable *X* from past data. But those data are generated by human activity within oppressive social and political structures and attendant ideologies, which make particular social identities highly relevant to a variety of outcomes.[18] And, in cases where someone's social identity is relevant to the predictive task, it is impossible to separate out the effects of discrimination and have a model suited for the predictive task.[19]

AI technologies are not only a mirror of the social world. They are also a powerful *shaper* of our institutions and social practices.[20] Hiring processes, for example, incorporate algorithmic decision-making to find candidates, sort applications, and conduct initial interviews. Algorithms used in hiring have been shown to reproduce societal biases. But increased data sharing and feedback loops between algorithmically mediated decisions, the data, and algorithms can also increase unjust disadvantage for members of oppressed or marginalized groups. [21] For example, the data that determines one's credit score often appears in a credit report, which about one-half of employers in the US use in the hiring process.[22] Biased data can influence both employment outcomes and outcomes that require access to credit, such as education and housing. And algorithmic decision-making can create new social identities that are the subject of discrimination, such as the algorithmically left out[23] or the commercially unprofitable.[24]

The moral concern about algorithmic bias is the concern that algorithmic decision-making can be *wrongfully discriminatory*.[25] Wrongful discrimination is unjustifiably differential treatment. To further flesh out this idea, I will borrow two standards for wrongful discrimination from US federal law: *disparate treatment* and *disparate impact*. Disparate treatment involves treating someone differently because of a social identity characteristic, in a manner that is inconsistent with their equal moral worth. The wrongness of the treatment is often understood to be a matter of the decision-maker's intentions: treatment is wrongfully discriminatory when a decision-maker is motivated by negative attitudes about members of that group.[26] However, here I follow Hellman (2011) in thinking that no intention is required to treat someone in a discriminatory way. Treatment can be disparate when someone is *demeaned*, or treated as having lesser moral worth, in virtue of a socially salient characteristic. Disparate impact, by contrast, does not locate discrimination in an interpersonal interaction, be it one where one party intends to discriminate, or where one party demeans another. Instead, disparate impact occurs when there are unjustified inequalities between marginalized or oppressed groups and privileged groups.[27] While the law is an imperfect guide to morality, these two standards capture important moral concerns about discrimination: how we treat people, and how our institutions tend to allocate benefits and burdens among different groups. AI can be discriminatory in both senses, as we will see.

Let's now turn to our examination to the process of developing and deploying an AI system for decision-making. I will focus on major points that bias can enter into AI systems. The first step to decide on the task that the system is to perform. Imagine that a company wants to invest in an algorithm to streamline the hiring process. They want this algorithm to identify the best job candidates from a pool. "Find good employees" is not a task that a machine can accomplish. The task must be the right kind of task, namely, a problem that can be solved by predicting the value of some target variable. And the target variable must be one that can be measured, and the model must be built, tested, and deployed using

available data. "Predict who would make the most sales in their first year on the job," is, by contrast, a task that an algorithm can accomplish.[28]

However, certain ways of formulating tasks can be discriminatory. Disparate impact is a particular worry regarding task choice, especially because it is not always obvious when a certain problem specification will lead to unjustifiable differential impacts on protected groups.[29] Disparate impact can come about because the task specification leads to differently accurate predictions for members of different groups. For example, Obermeyer et al. (2019) show that a common healthcare algorithm is less accurate in predicting Black patients' comorbidity score; that is, Black patients have a higher comorbidity score than white patients at the same level of assigned algorithmic score. They hypothesize that the bias arises because the algorithm does not predict a comorbidity score; instead, it predicts total medical expenditure in a year. But total medical expenditure is not a good operationalization of health need for Black patients because Black patients generate fewer healthcare costs at a given level of health.

The choice of task can also lead AI systems to be discriminatory, in the disparate treatment sense. Consider research by Latayna Sweeney (2013) that found significant discrimination in ad delivery by Google AdSense. Sweeney found that a greater percentage of ads containing "arrest" in the title were delivered for searches of names that are racially associated with being Black than for searches of names that are racially associated with being white. One plausible mechanism that produces this difference is the target metric—optimizing for clicks. By optimizing for clicks, Google and other ad systems create the conditions for discriminatory user behavior to determine ad relevance. By clicking ads with "arrest" in the title more often for names that are associated with being Black than for names that are associated with being white, users make ads with "arrest" in the title more relevant to Google searches for names associated with being Black. Thus, the task—coupled with data generated by the users—creates associations between Blackness and criminality. AI systems may not have intentions to discriminate or discriminatory attitudes. But an ad system that associates Black names with criminality is surely demeaning, qualifying as disparate treatment.

The next step of the AI pipeline is to gather data and extract features. Discrimination can arise here as well. One will often hear the adage "garbage in, garbage out" to describe the idea that one's model is only as good as the data that it is built on. One way that data can lead to discrimination is by being inaccurate. Measurement error includes oversampling or undersampling from a population or missing features that better predict outcomes for some group in the feature extraction process. Inaccuracy can lead to disparate impact because the algorithm will make more mistakes about one group. The remedy is usually to gather better data.

However, in some cases, the problem is not measurement error. In those cases, the bias is due to discrimination in the institutions and other social contexts that generate the data.[30] For example, Richardson et al. (2019) argue that police corruption and other unlawful practices have seriously degraded the quality of data in many places in the United States, making accurate feature choice to train predictive policing models incredibly difficult.[31] In such cases, collecting more and more accurate data will not address the problem of discrimination.

The next step in the process of building an AI system is to build a model using the data. Here is a third place where bias can enter, in the form of model bias. Models, especially

those learned through machine learning techniques, can learn biased statistical regularities from the human-generated data they are trained on. A major source of model bias is biased data, but other modeling decisions can introduce bias. Machine learning, for example, aims to find the optimally performing model for some task relative to a particular standard.

The choice of which metric to optimize for can have discriminatory impacts. For example, optimizing for predictive accuracy can worsen disparate impact discrimination. Predictive success is often achieved by using *proxy attributes*, or features that correlate with some socially salient characteristic, to the target of interest.[32] Even if proxy attributes have only a small correlation with a sensitive attribute, access to many such features will produce a classifier that implicitly uses sensitive attributes for prediction.[33] Prediction based on proxy attributes can thus induce a tradeoff between accuracy and the demand not to discriminate. Because social identities pervasively structure our social lives, reducing a reliance on proxy attributes can reduce the model's accuracy. But, if those predictions have negative, differential impact on protected groups, then we have discrimination-based reasons to reduce the reliance on those features in decision-making.[34]

The final step where bias may enter is when an AI system is used for decision-making. To use AI for decision-making, decision-makers must convert predictions to decisions. *Decision thresholds* are often used to convert a continuous prediction into a decision. A decision threshold is a function from a prediction to a decision. Decision thresholds are useful because they take a more complex prediction, often in the form of a probability distribution, and assign individuals into a "yes" or a "no" set, based on whether they are over or under the threshold.

Features of the context in which the algorithm is deployed may introduce bias. The decision-maker's goal may be to identify members of disadvantaged or marginalized groups in order to further disadvantage them.[35] The performance of the model may also decrease if the model is used in a context where the data is very different from the data that the model was trained on.[36] If the model is systematically less accurate for members of disadvantaged or marginalized groups, then the model can create disparate impact, as discussed in the context of healthcare above. Furthermore, systematic deployment of a model across many different contexts can create feedback loops, further entrenching disadvantage. For example, if employers tend to use a small number of hiring algorithms that disfavor older workers, future algorithms will not have much data about successful older workers, leading to even fewer that are hired.[37] This could also lead to disparate treatment, as being older becomes associated with being unemployable.

Accurately predicting an individual's risk of being caught committing a crime, and then making a bail decision on that basis, can further disparate impact discrimination. And, the decision-maker may have multiple goals, some of which are not represented in the target metric.[38] But, if a decision-maker is concerned about disparate impact discrimination, algorithmic predictions can still be very useful. In response to the accuracy-anti-discrimination tradeoff discussed earlier, some have suggested that some algorithmic predictions should be considered as diagnostic for the presence of injustice.[39] Prediction still guides action, in such cases, but it is more likely to point the decision-maker towards remedying background injustice that produces disparate impact discrimination, rather than bearing on the decision at hand. This allows decision-makers to use accurate predictions to achieve anti-discrimination goals, and thereby to mitigate the tradeoff between the two.

AI thus raises serious concerns about wrongful discrimination due to algorithmic bias that can enter at different points in the design and deployment of AI systems. Because AI is a powerful mirror and shaper of the social world, we urgently need to better understand the problem of algorithmic bias, and what to do about it. The rest of this chapter takes on that task.

A Problem of Fairness, Or A Problem of Justice?

Researchers, civil society activists, politicians, and businesses have generated a wide range of policy proposals, technical tools, and governance strategies to make algorithms fairer. For any of these, however, there is trenchant disagreement. Consider ProPublica's 2016 reporting on COMPAS, an algorithm that informed judicial decisions in the United States criminal justice system. They alleged that the risk assessment algorithm, which predicted a defendant's risk of future crime, was biased against African Americans. The developers of COMPAS, however, contended that the algorithm was indeed fair. What can explain these different judgments?

In this section, I will argue that research, governance, and public debate about algorithmic bias contain a key unarticulated disagreement over whether it is more pressing to govern AI to be *just* or *fair*.[40] "Justice" is a term that encapsulates a number of moral commitments about how society's key institutions ought to treat people. Fairness is one type of standard of justice.

There have been increasing demands for *justice* in AI. "Justice" refers to the moral standards that ought to structure major societal institutions, such as legal, political, and economic institutions.[41] These standards determine people's obligations, entitlements, opportunities, and burdens. In doing so, they unavoidably shape people's life trajectories, as well as their ideas about justice. One example is a country's laws about income tax. Tax law determines how much everyone must pay in income tax, which in turn determines the level of inequality in a country, how much individuals have to work to sustain themselves, incentives to work, and so on. The moral standards behind income tax law also determine people's ideas about the justness of different income tax regimes. If, for example, the tax regime is based on the idea that people deserve their pre-tax income, then citizens may not be willing to support higher income taxes.

Taking the perspective of justice allows us to better interrogate issues around AI and inequalities of material resources, opportunity, and power, individual autonomy and political self-determination, representation, and respect and recognition. In other words, it allows us to take a structural perspective on institutions and the moral standards they ought to live up to, but usually fail to.[42] We can ask questions of *distributive justice*: how basic rights, liberties, opportunities, and material resources ought to be distributed. Or we can ask questions about *productive justice*: how work ought to be organized, such that the burdens and benefits of work are justly apportioned. We can also ask questions of *corrective* justice—how should institutions right past wrongs, especially in cases of historic injustice.

Or of *racial and gender* justice—how structures perpetuate the exercise of power by some groups over others. Thinking about justice opens a wide swath of important structural questions (see the chapter on AI and structural injustice in this volume).

Standards of justice are grounded in a plurality of different types of moral reasons: reasons of equality, autonomy, gratitude, or dessert. One important category of such reasons are reasons of *fairness*, which embody a valuable type of equality. Fairness is about *respecting equal claims*, as well as respecting claims in proportion to their strength (Broome, 1991). Reasons of fairness are grounded in the moral equality of people. All else being equal, people have an equal claim to important resources to make their life go well, or to equality of respect and standing in their political community. Of course, all else may not be equal regarding material resources or other goods. Someone may have a claim to more of a good because of a prior agreement that establishes a greater claim, or for the reason that they deserve or need more of the good. In those cases, fairness requires respecting people's claims in proportion to the strength of those claims. Thus, fairness is a matter of *proportional equality*, or giving people their due relative to what is owed to them.

Fair decision procedures are those that treat like cases alike, in terms of people's claims.[43] If two people have a claim to the same level of worker's compensation based on injury, then a fair decision procedure gives them both that level of worker compensation. A fair outcome is a distribution of resources, material or not, that respects people's claims. And, alongside examining distributions, we can also interrogate whether decisions using AI satisfies people's equal claim to respect, or to not be discriminated against in the disparate treatment sense.

Fairness is a central moral concept in our thinking about justice and AI. However, it is not the only moral concern one might have. Some of the disagreement over algorithmic bias comes from a disagreement over whether to pursue fairness, or whether to pursue some of the other values of justice. Consider debates over the use of AI for predictive policing in contexts there are racial disparities in the past crime data. Let's say that those disparities come about due to different base rates in the crimes committed by members of each group.[44] In such contexts, calls for fair models often motivate calls for *accurate* predictive policing. Accurate predictive policing, one might claim, respects people's equal claims to protection from law enforcement. This can be done through predictive models that use accurate statistical generalizations about the actual populations the model is trained on (call this a "bare statistic") to make predictions. But those bare statistics may only be true of the arbitrary populations the model was trained on; they may not be *projectable*, or true in new populations. One important reason that bare statistics may not be projectable is that statistical generalizations are true against a background of particular social structures and practices, i.e., certain economic and social determinants.[45] A concern for racial justice, by contrast, more often recognizes that these statistics are not projectable. Calls for racial justice in policing are often calls to ban algorithmically driven predictive policing, or to use AI for other interventions to reduce the difference in base rates.[46] Such calls are concerned that a focus on accurate prediction of crime may, perversely, increase the unjust subordination of one group by keeping in place the social conditions that underwrite the generalization and entrenching the pattern by using it as an evidential basis for policy. Calls for racial justice are calls to increase the justice of the background structures that produce crime, rather than targeting an equal application of current legal standards.

The disagreement about fairness and other values of justice inflects a number of disagreements over algorithmic bias. In the next section, I will examine one such disagreement in more detail: the disagreement over fairness metrics.

Illustration: Fairness Metrics

There has been an explosion of research in computer science on algorithmic fairness, e.g., how to develop less discriminatory algorithms from a technical standpoint. *Statistical* approaches are post-processing techniques that take a learned model and aim to measure and mitigate discrimination based on observable inequalities between groups. *Causal* approaches are often taken to correct shortcomings in statistical approaches.[47] I will not attempt an exhaustive survey of this fast-paced, interdisciplinary area of research. Instead, I will use the example of statistical fairness criteria to illustrate the divide between proponents of fairness and proponents of other values of justice.[48]

Different statistical criteria for fairness have been proposed. These fairness criteria each formalize a notion of fair prediction as prediction where the output prediction is non-discriminatory or does not depend on individuals' social identities. Thus, most fairness criteria are expressible in terms of a relationship of (conditional) independence between the predictor R, the sensitive attribute A, and the target Y (i.e., the algorithm's task).[49]

Statistical fairness criteria exploded onto public consciousness because of ProPublica's (2016) allegation that COMPAS, an algorithm that predicts a defendant's risk of future crime, was biased against African Americans. ProPublica made this allegation based on a comparison of the predicted risk score against later data about who committed crimes. Their analysis found that the algorithm wrongly labeled Black defendants as high risk at about twice the rate as white defendants; conversely, it wrongly labeled white defendants as low risk at almost twice the rate as Black defendants. They concluded that the algorithm was discriminatory. That conclusion assumes that *false-positive or false-negative equality* is the best criterion for detecting algorithmic bias. These criteria measure the proportion of false positives or false negatives within each group. Such false-positive or false-negative equality metrics are motivated by the idea that people who are the same with respect to outcomes ought to be treated the same in terms of having true (or false) decisions made about them at similar rates. More formally, conditional on the target variable Y, the score R should be independent of protected attribute A.[50]

The developers of COMPAS, however, contended that false-positive or false-negative equality is not the best measure of whether the algorithm is accurate. This point has been compellingly argued by a number of philosophers, computer scientists, and social scientists.[51] Here is one example argument as to why a difference in false-positive or false-negative rates is not always an indicator of unfairness. Recall that AI systems often employ a threshold to convert a prediction into a decision. There would be fewer false positives for individuals that are predicted to be far from the threshold (more colloquially, the clear yes or a clear no), and more false positives and negatives closer to the threshold. If members of one group cluster closer to the threshold, and members of another group are on the high and lower ends of the thresholds, then there will be more false positives and negatives

for the former group. This difference in false-positives and false-negative rates between groups is not in and of itself unfair; to put it another way, the difference is not unfair in all circumstances.[52] So, if the best criterion is one that an unbiased system must always fulfill, then false-positive and false-negative equality is not such a criterion.

Many computer scientists, social scientists, and philosophers favor *calibration* as the best metric for fairness. A well-calibrated system is one in which, conditional on a particular score *R*, an individual's group membership *A* and the outcome variable *Y* are independent. The intuitive idea behind calibration measures is that the standards that the decision system uses work equally well to identify who meets that standard, independent of their group membership. It is often taken to formalize a notion of "fair testing" or "fair application of a standard": a test or application of a standard is fair to the extent that a decision is an equally good predictor of the actual outcome of interest. For example, a calibrated university entrance exam is one whose classification of students to admit and not admit based on their test scores is an equally good predictor of their success at university across different demographic groups.

Calibration thus seems to better embody a requirement of fairness, understood as treating like cases alike. Consider the example of a miscalibrated entrance exam, where a score of 90, say, is associated with an 85 percent chance of success at university for members of group 1, but only a 65 percent chance of success for members of group 2. If a decision-maker admitted all students with a score of 90 percent or more, members of group 1 have a 15 percent chance of a false positive, whereas members of group 2 have a 35 thirty-five chance of getting a false positive. Because members of group 2 have a systematically higher chance of being subjected to false positives, the system is unfair.[53]

Calibration is often satisfied without any explicit intervention, in cases where a sensitive attribute can be predicted from other attributes.[54] This fact is not surprising: if there is enough data about individuals from different groups that can be used to build an accurate predictor for a target that is correlated with group membership, then the scores that the model assigns to individuals are equally informative for the outcome. Interventions to ensure that systems are calibrated aim to produce accurate models, holding fixed background social structures. However, as discussed above, accurate modeling within unjust institutions can replicate and further injustice. Thus, while calibration may be a good metric to promote fairness, relying on calibration alone to mitigate algorithmic bias can set back other values of justice.

If one is concerned about these other values of justice, then unequal false positive or negative rates may be a better metric. That is because differential rates of false positives may be evidence of injustice, even if they are not unjust or unfair in themselves. For example, if more members of a privileged group tend to cluster away from a threshold, it could be because of past injustice in that group's access to resources. Or more false positives may have a disproportionate impact on the welfare of members of a marginalized group. However, it is important to be clear about the claim here. A difference in false positive or false negative rates is evidence of injustice, not unjust itself. Thus, a decision maker does not have reason to reduce false positive or false negative rates, unless they have evidence that doing so is a good means to increase welfare or reduce injustice. For example, one could reduce the disparity between white and Black individuals in COMPAS by prosecuting more Black individuals who are low risk. But, to do so would be unjust.

Justice Over Fairness

I've argued that different technical interventions can promote different values of justice, such as fairness and racial justice. Furthermore, I've also argued that there can be tradeoffs between promoting fairness and other values of justice. These lead us to an important moral question: which of those values should be prioritized?

There is not always a tradeoff between fairness and other values of justice, however. That is because fairness has no value in situations of serious and pervasive injustice. Fairness depends on other standards of justice because these other standards are necessary to determine which cases are alike, from a moral perspective. On its own, the value of equality does not speak to the question of which cases are alike; to put it another way, standards of fairness are "empty" unless other norms of justice determine which properties of agents count as relevant to the decision at hand.[55] Ought individuals be provided with unemployment benefits because they cannot work but are owed the means to live a decent life, because they have made valuable contributions to society but can't find a job at the moment, or just in virtue of being a member of the community? These different ideas of what justice requires back different claims; it is then a further question of fairness whether institutions respect those claims, *given what other standards of justice have fixed them to be*.

This point leads us to my claim of this section: If the contexts in which AI is designed and deployed are seriously unjust, then one must prioritize other values of justice over fairness.

The first argument for this claim is that without just institutions, fairness is of little moral worth. The value of fairness depends on the existence of just institutions in the background. Institutions determine at least some of people's claims, e.g., which individuals count as relevantly similar, as discussed above. However, mere equality of treatment, in the face of any possible set of claims, is not morally valuable. If the standards that determine individuals' claims are unjust, then applying rules in an even-handed manner does not have moral value. Say that I make a rule that the first 10 children to arrive at a birthday party get a slice of cake, and the rest don't get any cake. I faithfully implement this rule at the birthday party, giving the first 10 children a slice of cake, and the rest no cake. For the children who didn't get any cake, is it morally valuable that I fairly applied the rule? Arguably, not. Implementing a rule consistently or impartially is not in itself valuable. We value consistent rule application when the rules are just in their distribution of benefits and harms. In other words, fairness is an important value of justice because people have *legitimate* claims, that ought to be respected. Fairness then ensures that those rules are used consistently and impartially in decision-making, respecting the claims of each.

One reason, then, to prioritize other values of justice is that without justice, there is no fairness. The second reason to prioritize other values of justice is one that I have already discussed in some detail throughout. The reason is that fair decision-making can compound injustice. Imagine that there is enough good housing for everyone in a city at affordable rates to rent or own, but that higher-quality housing is more expensive, and lower-quality housing is less expensive. Further, imagine that the political and social institutions of this society hold the norm that everyone deserves quality housing, and that housing of different quality should be distributed according to willingness to pay. Landlords decide between everyone who can afford the housing by lottery. However, this society is also marked by

historic and current access to credit along racial and gender lines. The landlord's decision lottery is fair, as it respects everyone's equal claims to a house. But it compounds injustice, in that ability to secure a mortgage or loan for a rent shortfall on fair terms is determined in party by race and gender. If a fair decision procedure relies on inputs that are the outputs of historic or current structural injustice, then the decision procedure will compound injustice. Because not compounding injustice is more important than fairness, we should focus on whether AI compounds injustice, if there is a tradeoff between that value and fairness.

Policy Recommendations for Just AI

I want to close by discussing policy strategies to govern AI for justice, including fairness. These policy strategies come out of the preceding discussion of algorithmic bias, justice, and fairness. They are not intended as a complete survey of the governance options.

Take a values-first approach to bias interventions

It can be tempting to introduce AI into existing decision processes without reflection on the new moral problems that AI raises, or how it might exacerbate existing problems. Furthermore, companies and others who benefit from easily available data have tried to inculcate the attitude that the use of any data is fair, as long as it is predictively useful.[56] This can further blind us to the ways in which algorithmic choices are value laden. An especially gaping gap in algorithmic governance is a moral examination of the predictive task, as interventions to measure and reduce algorithmic bias tend to focus on modeling choices and model-based interventions, as well as, to a lesser extent, the quality of the data.

The first governance suggestion is to take a values-first approach to algorithmic bias. By "values-first," I mean that clear statements of values and their tradeoffs ought to determine the choice of a policy goal for AI regulation or the measurement and mitigation of bias in a particular algorithmic system. Sometimes, in combatting bias, it seems as if the goal is clear—say, to increase diversity. But we can better understand what the goal is, and what interventions can achieve it, if we recognize the values behind increasing diversity. In hiring, for example, a decision-maker may be concerned that equality of opportunity is violated because qualified candidates from one race are not noticed as often by recruiters (perhaps the company is racially homogenous, and employees often recommend their friends for jobs at the company). If the value is equality of opportunity, then steps to increase diversity in sourcing—such as auditing job ads for potential biased language—may be called for. In university admissions, by contrast, a decision-maker may be concerned that students from disadvantaged backgrounds are subject to further disadvantage because they did not have the same opportunities to build their resume that other students had. Here, an affirmative action system may be warranted by the value of reducing educational injustice.

A values-first approach is especially important for algorithmic bias because different values of justice support different standards to measure bias in the system and support different interventions. Regulatory and governance efforts must be clear about what the

important moral values are in that domain, and how the algorithm's task, data, modeling choices, and design connect with those values. Take risk assessment tools in criminal justice. What kind of risk ought they to be predicting, in contexts with historical injustice and current structural injustice? Say that legislators aim to avoid racial injustice and keep people safe from violent crime. These two values support decision-making tools to predict the risk of violent crime arrest. Of course, there may not be enough data to predict violent crime, as it is rarer than other kinds of crime; in that case, decision-makers should not substitute another target that is easier to predict. AI development and deployment should be driven by values, not by data availability.[57]

De-couple decision processes

The second governance suggestion is to de-couple decision processes. This point holds across decision process, and for a single type of decision, such as loans or hiring. As argued above, AI can compound injustice across decision processes. Furthermore, because AI systems can operate at a much greater scale than well-coordinated human decision makers, and they can create salient social identities that may be the target of injustice, there are strong reasons of justice to avoid using the same AI system for a large class of decisions.[58]

Governance efforts can target AI-generated decision aids that are commonly used across different types of decisions. Credit reports, for example, are used to make decisions not only about loans, but also about jobs, housing, and insurance, even though there is not much evidence that credit reports are a good predictor of productivity.[59] But, because personal data about individuals is cheap and easily available, this move will be of limited utility on its own. Governance efforts will also need to reduce the sale and use of personal data.[60]

A concern about this strategy is that it reduces the accuracy of decisions by limiting the input data for building AI systems. In many cases, such as the use of credit reports in hiring, data are used to make efficient decisions that are better than choosing at random, not highly accurate decisions. Furthermore, governance strategies can promote accuracy in the long term. One such governance strategy promotes the use of lotteries early in a decision process, say, a lottery among interested law students to hire summer interns at a law firm, to avoid compounding injustice. After a certain amount of time, employers will have enough data about the actual job performance of those interns to make a data-driven decision.[61] Another kind of governance strategy introduces multiple decision processes based on different criteria. A diversity of criteria can avoid over-indexing on a single set of criteria, which are likely to arbitrarily privilege those who have been the recipients of past advantage.[62]

Model structural injustice

A third governance suggestion is to include information about structural injustice in the decision process, especially for decision-makers that ought to advance justice.[63] How computer and social scientists model the social world depends on their assumptions about the prevalence and severity of racial injustice.[64] If racial injustice is prevalent, then scientists

ought to be cautious about including proxy attributes for race, limiting what they can predict. Scientists ought to make these empirical assumptions explicit, so that decision-makers are in a better position to judge whether the algorithm's assumptions hold.

Furthermore, decision-makers that want to promote gender and racial justice need to understand where and how they can intervene, and what the downstream effects of using an AI system will be. Predictive algorithms tend to take institutional background structures as given, which can sideline possibilities for intervening on those institutions to promote more just outcomes.[65] And, the moral permissibility of using a particular decision system depends on its long-term impacts in a particular context. Without modeling the dynamics of that social context, the decision-maker does not have a proper evidence base to judge whether to deploy the AI system.[66]

Better data

The fourth suggestion is more regulation to improve data quality. Data is often taken as given and as fact; furthermore, there is a tendency to use data that is easily available. But, as we saw, biased data is one of the key causes of model bias. Increasing data quality will produce more accurate and robust models.

There are a number of scientific, social, and political interventions required to produce better data. Some regulations that would improve data quality are more scientific or technical.[67] Other regulations would instead correct incentives in the data economy that produce poor or limited data. For example, as Véliz discusses in this volume, data brokers have incentives to collect massive quantities of data cheaply, and a disincentive to ensure that any piece of information is accurate. Banning targeting advertising would address data quality issues. In its place, governments should consider funding public organizations to create accurate and inclusive data sets. They should also mandate data sharing by private companies with researchers and auditors.

Replace decision thresholds with more (weighted) lotteries

The previous governance strategies focused on other values of justice. The final suggestion focuses on fairness. If one is concerned about designing for fairness, then one should not use decision thresholds. Instead, fairness requires more decision-making by lottery. Alongside calibration, randomness offers another technical lever to increase fairness.

It is important to be clear upfront about what these arguments purport to show. They show that decision-makers have reasons of fairness to use lotteries instead of decision thresholds. They do not show that decision makers have decisive reason to introduce more randomness into algorithmic decision-making. There may be other, weightier reasons to use decision thresholds.

If decision-makers know what an individual's claims are, as well as their strength, then they should use weighted lotteries, rather than decision thresholds. The argument for this claim relies on the definition of fairness as the satisfaction of claims in proportion to their relative strength. Say that Carl has lent me $100, and Dani has lent me $200, but I only have $60 to pay them back. In such a case, fairness requires that I ought to give Dani $40 and

Carl $20 because her claim is twice as strong as Carl's. In the case of divisible goods like money, the good should be allocated proportionally, where the proportions are determined by the relative strength of people's claims.

And, when goods are indivisible, they should also be allocated proportional to the strength of someone's claim, by weighed lotter. Let's now change the example to a kidney transfer. Dani is four times as sick as Carl, and thus has a claim to the kidney that is four times as strong as Carl's. The fairest way to allocate the kidney is by a weighted lottery, where the weights reflect the proportional strengths of various claims. A kidney lottery, for example, should be set up so that Dani is four times more likely to win the lottery as Carl.

This argument may strike you as objectionable—shouldn't the person with the strongest claim get the kidney? At first glance, a policy that always distributes a good to someone with the strongest claim seems to be the fairest policy.[68] Such a challenge seems more challenging for a single decision, say, whether to give the kidney to Carl or Dani.

However, such a policy ignores the weaker claims of others. The moral strength of this point is more easily appreciated when one zooms out to reason about the fairness of an allocation procedure for a population over time. Imagine that this kidney allocation lottery was used across a country to allocate donated kidneys to prospective patients. Fairness requires that the decision-maker can give the losers of an allocation procedure—those who don't win the kidney lottery, say—a reason why their claims were respected. And, in the case of the kidney allocation, the reason must be that they had a real chance at getting a kidney.[69] It does not seem reasonable to ask them to completely sacrifice their claims to the someone with a stronger claim, which would be required by an allocation procedure that always allocated indivisible goods to those with the strongest claims. A weighted lottery, by contrast, respects everyone's claims, as individuals have a chance of getting the good that is proportional to the strength of their claim. Thus, fairness requires weighted lotteries for indivisible goods—and, more generally, that claims be satisfied in proportion to their strength.

This point raises a serious challenge to the allocative fairness of most algorithms. For most algorithms, those below the decision threshold have some claim to the desirable outcome (or to avoid an undesirable one). However, anyone below the decision threshold has no *ex ante* real chance at the outcome. So, any algorithm that uses a decision threshold to separate individuals with greater and lesser claims is unfair.

The argument for weighted lotteries assumes that individuals have well-established claims of differing strengths, and that decision-makers can gather enough information about those claims to design a weighted lottery. In situations of injustice, it may be that individuals have an equal claim to a good, rather than claims of differing strength. For example, in a society where people have unjust differential access to educational opportunities and material resources, it may be that better qualified individuals do not have a greater claim to the job. So, the initial lottery for all qualified individuals would be fairer, rather than allocating the job to the most qualified individual. Furthermore, individuals may have claims of differing strengths, but decision-makers may not be able to gather enough information to identify which individuals have stronger claims. In such cases, a lottery with equal weights would be fairest, as each individual has the same *ex ante* chance of their claim being disregarded, in light of what the decision-maker knows.

Conclusion

"Fairness" is an ambiguous and messy term, one that is central to politics but can be more obfuscating than clarifying. This chapter distinguished between fairness and other reasons of justice, and explained disagreements over how to address algorithmic bias as disagreements over whether to prioritize fairness or those other reasons of justice. One major lesson of the chapter is that more accurate decision-making can contribute to injustice. However, less accurate decision-making is not the solution, as it may not advance justice, and can come at too high a cost to other values. The chapter ended by promoting a number of governance strategies to promote fairness, such as reducing the use of decision thresholds, and to reduce the injustice that can arise from AI systems that exploit predictively powerful correlations that increase inequalities or otherwise entrench disadvantage, such as modeling structural injustice and better data.

Notes

1. Danks and Fazelpour (2021a).
2. I take "discrimination" to be a morally neutral term. I may discriminate between wrestlers by assigning them to groups based on their weight or discriminate between students by organizing them by last name. These examples motivate that it is not always wrong to draw distinctions among people based on certain properties that they have (Hellman, 2011). For purported instances of differential outcomes based on group membership, it is important first to ask whether such discrimination is wrongful.
3. Definition based on Johnson (2020, p. 9942).
4. Buolamwini and Gebru (2018).
5. Angwin et al. (2016).
6. Ali et al. (2019).
7. Guardian Editorial (August 11, 2020).
8. Apple, for example, faced legal scrutiny after the Apple Card granted lower credit limits to women than men (AI Now, 2019).
9. Dotan (2020); Johnson (forthcoming).
10. See Fischer et al. (1996) on how policies have widened the wealth gap in America, or Mills (1999) on white supremacy and the so-called racial contract, where social institutions are set up to promote white equality and interests and subordinate people of color.
11. Shoam et al. (2018) found that 80 percent of AI professors were men.
12. For example, Brooke (2021) finds that gender bias determines knowledge sharing and recognition on Stack Overflow.
13. Crawford et al. (2019).
14. Philosophers of science have argued that the potential impacts of scientific choices mean that scientists ought to consider social values throughout the scientific process (Douglas, 2007). This literature focuses on the costs of false positives and false negatives, in line with the algorithmic fairness literature (Section 4).
15. Benjamin (2019); Gabriel (forthcoming).

16. This insight is a central and longstanding one in feminist philosophy and the philosophy of race. As James Baldwin (1998, p. 723, from Benjamin 2019, p. 5) observed: "The great force of history comes from the fact that we carry it within us, are unconsciously controlled by it in many ways, and history is literally *present* in all that we do."
17. For example, Eubanks (2018), Noble (2018), and Mayson (2019).
18. As Frye (1983, p.19) says: "It is not accurate to say that what is going on in cases of sexism is that distinctions are made on the basis of sex when sex is irrelevant; what is wrong in cases of sexism is, in the first place, that sex is relevant; and then that the making of distinctions on the basis of sex reinforces the patterns which make it relevant."
19. One strategy is to remove the social identity variables as training data for the learning process or inputs to the prediction model. But, if the social identity is a significant predictor for the target, then many other predictively useful inputs will be correlated with social identity as well (the so-called "proxy problem") (Corbett-Davies & Goel, 2018).
20. Gabriel (forthcoming); Noble (2018).
21. Gandy (2009); Hellman (forthcoming), and Herzog (2021).
22. Kiviat (2019). It is often used as a proxy for responsibility (O'Neill, 2016, p. 147).
23. O'Neill (2016).
24. Fourcade and Healy (2013) discuss the ways in which companies can use AI and big data to predict who will be a profitable customer, creating new, economically and socially salient categories of the profitable and unprofitable.
25. Hellman (2011, pp. 2–4).
26. Alexander (1992). For criticism of such accounts, see Hellman (2011), and Lippert-Rasmussen (2013, Chapter 4).
27. Selmi (2013, Chapter 12). A theory of disparate impact must specify when inequalities qualify as discrimination, lest, for example, it condemns affirmative action, or harmless inequalities that happen to come about through group member's choices. More analysis is needed to state when an inequality is unjustified. But, since disparate impact is not the focus of this chapter, I will set such issues aside.
28. Passi and Barocas (2019).
29. Barocas and Selbst (2016).
30. Mayson (2019).
31. See also Knox, Lowe, and Mummolo (2020), as well as Gabler et al. (2020) for pushback.
32. Dwork et al. (2012); Johnson (2021). Discrimination that arises from decision-making based on proxy attributes has long been a concern to philosophy of discrimination and the law (see, e.g., Alexander, 1992).
33. Barocas, Hardt, and Narayanan (2019).
34. Johnson (2021).
35. Eubanks (2018) chronicles how AI can be used by government officials to further disadvantage the poor.
36. This is the so-called problem of external validity: under what conditions does a model predicts or explains well in new contexts, and when should decision-makers be confident that the model will do so? (Rodrick, 2009)
37. Herzog (2021).
38. Kleinberg et al. (2018) call this "omitted payoff bias."
39. Mayson (2019).
40. This diagnosis is a more general version of Barocas and Selbst's (2016) argument that attempts to use United States anti-discrimination law to tackle algorithmic bias suffer

from an ambiguity in the law, as to whether anti-discrimination is an anti-classification or an anti-subordination project.

41. Rawls (1999).
42. Le Bui and Noble (2020) call for a moral framework of justice to be integrated into research and governance on AI. They, following Mills (2017), push for an explicitly non-liberal moral framework of justice.
43. Hart (2012). Zimmermann and Lee-Stronach (2021) call this the "Like Cases Maxim."
44. As previously discussed, disparities in law enforcement data often come about through racially biased policing (Richardson et al., 2019; Mayson, 2019). This chapter does not assume that any actual disparities in past crime data are due to a difference in base rates.
45. Munton (2019).
46. Mayson (2019).
47. Kusner and Loftus (2020).
48. Statistical fairness criteria are also worth focusing on because they are, arguably, still the standard approach in computer science and in industry to quantify the amount of bias in an algorithm (Danks and Fazelpour, 2021a).
49. Barocas, Selbst, and Narayanan (2019, Chapter 2).
50. Where R is a binary classifier and *a* and *b* are two groups, this can be expressed in terms of the conditional probability that $P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$ for false positive equality, and $P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$ for false negative equality.
51. E.g., Corbett-Davies & Goel (2018), Long (2022), and Hedden (2021).
52. For similar arguments, see Hedden 2021.
53. See Corbett-Davies & Goel (2018), Long (2022), and Hedden (2021) for arguments that calibration is a requirement of fairness.
54. And calibration does not ensure that design choices—the target goal, or the decision function to translate predictions into scores—are just. Say that a bank wishes to discriminate against Black loan applicants and knows that Black applicants live in zip codes with relatively high default rates, and that white and Black applicants have similar default rates within each zip code. The bank could develop a calibrated algorithmic system that predicted default rates by zip code alone. But, one could build a calibrated alternative system that uses more information and divides individuals into finer risk buckets (Corbett-Davies & Goel, 2018, p. 16).
55. Hart (2012); Westen (1982).
56. Zuboff (2019).
57. Mayson (2019, p. 2269).
58. Creel and Hellman (2022).
59. Weaver (2015).
60. See Veliz's contribution to this volume
61. Hu and Chen (2017).
62. Fishkin (2013).
63. Zimmermann and Lee-Stronach (2021) and Mayson (2019). See also Gabriel (forthcoming), Herzog (2021), and Ferretti (2021) for arguments that the duty to advance justice applies to private companies developing and deploying technology.
64. Hu (2021).
65. Zimmermann and Lee-Stronach (2021).
66. Danks and Fazelpour (2021b).

67. See Gebru et al. (2018) on data sheets for data sets.
68. Hooker (2013).
69. Spiekermann (2021).

References

Alexander, L. (1992). What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review 141* (1), 149.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. https://arxiv.org/pdf/1904.02095.pdf.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. There's software used across the country to predict future criminals. And it's biased against Blacks." *Pro Publica*, May 23.

Baldwin, J. 1998. *James Baldwin: Collected Essays*. Library of America.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review 104* , 671.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. https://fairmlbook.org/.

Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research 81* , 1–15.

Brooke, S. J. (2021). Trouble in programmer's paradise: Gender-biases in sharing and recognising technical knowledge on Stack Overflow. *Information, Communication & Society 24* (14), 2091–2112. https://doi.org/10.1080/1369118X.2021.1962943.

Broome, J. (1991). Fairness. *Proceedings of the Aristotelian Society 91* , 87–101.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A., Raji, D., Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). *AI Now 2019 Report*. AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html.

Corbett-Davis, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023v2 [cs.CY]*.

Creel, K., & Hellman, D. (2022). The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*: 1–18. doi:10.1017/can.2022.3

Danks, D., & Fazelpour, S. (2021a). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*. https://doi.org/10.1111/phc3.12760.

Danks, D., & Fazelpour, S. (2021b). Algorithmic fairness and the situated dynamics of justice. *Canadian Journal of Philosophy*: 1–17. DOI: https://doi.org/10.1017/can.2021.24.

Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese 198* , 11081–11101. https://doi.org/10.1007/s11229-020-02773-2

Douglas, H. E. (2007). Rejecting the ideal of value-free science. In H. Kincaid, J. Dupré, and A. Wiley (Eds.), Value Free Science? Ideals and Illusions. Oxford University Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *arXiv:1104.3913 [cs.CC]*.

Eubanks, V. (2018). *Automating inequality*. MacMillan Publishers.

Ferretti, T. (2021). An institutionalist approach to AI ethics: Justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics*. https://doi.org/10.1515/mopp-2020-0056

Fischer, C., Hout, M., Sánchez Jankowski, M., Lucas, S. R., Swidler, A., & Voss, K. (1996). *Inequality by Design: Cracking the Bell Curve Myth*. Princeton University Press.

Fishkin, J. (2013). *Bottlenecks: A new theory of equality of opportunity*.

Fourcade, M., & Healy, K. (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society 38* (8), 559–572. https://doi.org/10.1016/j.aos.2013.11.002.

Frye, M. (1983). *The Politics of Reality: Essays in Feminist Theory*. Crossing Press.

Gabriel, I. (forthcoming). Towards a theory of justice for artificial intelligence. *Daedalus*. *arXiv:2110.14419 [cs.CY]*.

Gaebler, J., Cai, W., Basse, G., Shroff, R., Goel, S., & Hill, J. (2020). Deconstructing claims of post-treatment bias in observational studies of discrimination. *arXiv:2006.12460 [stat.ME]*.

Gandy, O. (2009). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Ashgate Publishing Company.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H. Daumeé III, & Crawford, K. (2018). Datasheets for datasets. *arXiv:1803.09010*.

Guardian Editorial. (2020). The Guardian view on A-level algorithms: Failing the test of fairness. *The Guardian*, August 11. https://www.theguardian.com/commentisfree/2020/aug/11/the-guardian-view-on-a-level-algorithms-failing-the-test-of-fairness.

Hart, H. L. A. (2012). *The concept of law*. 3rd edn. Oxford University Press.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs 49* (2): 209–231

Hellman, D. (2011). *When is discrimination wrong?* Harvard University Press.

Hellman, D. (forthcoming). Big data and compounding injustice. *Journal of Moral Philosophy*.

Herzog, L. (2021). Algorithmic Decision-Making, Ambiguity Tolerance and the Question of Meaning. Deutsche Zeitschrift fur Philosophie *69*(2), 197–213.

Hooker, B. (2013). Fairness. *Ethical Theory and Moral Practice 8* (4), 329–352.

Hu, L., & Chen, Y. (2017). Fairness at equilibrium in the labor market. *FATML*. arXiv:1707.01590 [cs.GT].

Hu, L. (2021). Race, policing, and the limits of social science. *The Boston Review*, May 6. https://bostonreview.net/articles/race-policing-and-the-limits-of-social-science/.

Johnson. G. (2020). Algorithmic bias: On the implicit bias of social technology. *Synthese 198*, 9941–9961.

Johnson, G. (forthcoming). Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*. https://www.gmjohnson.com/uploads/5/2/5/1/52514005/are_algorithms_value_free.pdf

Kiviat, B. (2019). The art of deciding with data: Evidence from how employers translate credit reports into hiring decisions. *Socio-Economic Review 17* (2), 283–309.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics 133* (1): 237–293.

Knox, D., Loew, W., & Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review 114* (3), 619–637. https://doi.org/10.1017/S0003055420000039.

Kusner, M., & Loftus, J. (2020). The long road to fairer algorithm. *Nature 578* , 34–37.

Le Bui, M., & Noble, S. (2020). We're missing a moral framework of justice in artificial intelligence. *The Oxford Handbook of Ethics of AI*, Dubber, M., F. Pasquale, and S. Das (eds). Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780190067397.013.9

Lippert-Rasmussen, K. (2013). *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford University Press.

Long, R. (2022). Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy 19* (1): 49–78. https://doi.org/10.1163/17455243-20213439.

Mayson, S. (2019). Bias in, bias out. *Yale Law Journal 128* (8), 2122–2473.

Mills, C. (1999). *The racial contract*. Cornell University Press.

Mills, C. (2017). *Black rights/white wrongs: The critique of racial liberalism*. Oxford University Press.

Munton, J. (2019). Beyond accuracy: Epistemic flaws with statistical generalizations. *Philosophical Issues 29* (1), 228–240.

Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

O'Neill, C. (2016). *Weapons of math destruction*. Penguin.

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, Atlanta, GA, USA.

Richardson, R., Schulz, J. M., & Crawford. K. (2019). Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review 94* (192): 193–233.

Rodrik, D. (2009). The new development economics: We shall experiment, but how shall we learn? In Cohen, W. and Easterly, J. (Eds.), *What works in development?* (pp. 24–48). The Brookings Institute.

Rawls, J. (1999). *A theory of justice*. Revised edn. Harvard University Press.

Shoham, Yoav, Perrault, Raymond, Brynjolfsson, Erik, Clark, Jack, Manyika, James, Niebles, Juan Carlos, Lyons, Terah, Etchemendy, John, Grosz, Barbara, & Bauer, Zoe. (2018). The AI index 2018 annual report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.

Selmi, M. (2013). Indirect discrimination and the anti-discrimination mandate. In D. Hellman and S. Moreau (Eds.), *Philosophical foundations of discrimination law* (pp. 250–268). Oxford University Press.

Sweeney, L. (2013). Discrimination in online ad delivery. *arXiv:1301.6822v1 [cs.IR]*.

Spiekermann, K. (2021). Good reasons for losers: Lottery justification and social risk. *Economics and Philosophy*. *38*(1), 108–131. doi:10.1017/S0266267121000043

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 6464* , 447–453.

Westen, P. (1982). The empty idea of equality. *Harvard Law Review*, 95(3): 537–596.

Weaver, A. (2015). Is credit status a good signal of productivity? *Industrial and Labor Relations Review 68* : 742–770.

Zimmermann, A., & Lee-Stronach, C. (2021). Proceed with caution. *Canadian Journal of Philosophy*. 1–20. doi:10.1017/can.2021.17

Zuboff, S. (2019). *Surveillance capitalism*. Profile Books.

CHAPTER 7

GOVERNING PRIVACY

CARISSA VELIZ

This chapter explores what is privacy, why is it important, for whom it is important, and how we can better protect it. First, I offer what I call the hybrid account of privacy, according to which having privacy amounts to being personally unaccessed, and enjoying the right to privacy amounts to not losing control involuntarily over one's personal information or sensorial space. I then propose an explanation of why privacy is important: because it shields citizens from harms that arise from exposure to others. Next, I explore for whom privacy is important: for individuals and for collectives. Finally, I sketch some ideas regarding how we can better protect privacy in the context of AI. I will argue for data minimization, storage limitation, and banning the trade in personal data. I end the chapter with some thoughts on the role of consent.

What Is Privacy and the Right to Privacy?

In the digital age, you are losing privacy every day. As you walk the streets of your city, cameras capture your face and may even employ facial and emotional recognition on it. Your phone is being tracked for commercial and security purposes. Your browsing history, your purchasing records, and even your health data are being sold in the data economy. Governments and thousands of corporations may know more about you than most people you have a personal connection to.

Philosophical accounts of privacy can broadly be divided into access and control theories.[1] Most scholars who defend access theories of privacy define privacy as a matter of information being inaccessible or of limited access.[2] According to such views, you lose privacy when your personal information (or some other element that privacy is supposed to protect) becomes accessible to others. In contrast, according to control theories, you lose privacy when you lose control over your personal information (or some other element that privacy is supposed to protect). The philosopher Andrei Marmor, for example, has argued that "the underlying interest protected by the right to privacy is the interest in having a reasonable measure of control over ways you present yourself to others" (Marmor, 2015).[3,4]

The most important advantage of access theories is that they capture the intuition that sometimes we lose privacy without losing control. For example, if I tell something private to someone, I lose some privacy with respect to that person without ever having lost control over that information (we can further assume that that person is isolated in such a way that they cannot disseminate that information). Conversely, we can sometimes lose control over our personal information without losing privacy. For instance, if I forget my diary at a friend's house, but she decides not to read it.

In turn, the most important advantage of control theories is that they capture the intuition that it is wrong of others to make us lose control over our personal information, even when they do not access that information. For example, it would be wrong of my friend to steal my diary and store it, just in case she might feel like reading it in the future. Let's suppose that, unbeknownst to my friend, my diary is written in code such that she could not gain access to its content even if she wanted to. Even then, it seems like I have a privacy claim against my friend that access theories cannot capture.

An adequate theory of privacy must incorporate both access and control elements. I argue that privacy itself and losses thereof are better explained by an access theory, while the right to privacy and violations thereof are better explained by appealing to control. I will call the sum of both theories *the hybrid account of privacy*. An adequate access theory is a descriptive account that can help us answer questions related to when we have lost privacy. An adequate control theory is a normative account that can help us answer questions related to when our right to privacy has been violated. We need both to make sense of privacy.

The hybrid account of privacy

Privacy is the quality of having one's personal information and one's personal "sensorial space" unaccessed. In other words, you have privacy with respect to another person to the degree that the other person has not accessed your personal information and your personal space—that is, to the degree that they do not know anything personal about you and that they cannot see, touch, or hear you in contexts in which people typically do not want others' attention. Personal information is the kind of information about oneself that people in a certain society have reason not to want anyone, other than themselves (and perhaps a very limited number of other people chosen by them), to know about (e.g., because it makes them vulnerable to others' abuses of power).

One of the differences between this access account of privacy and that of others is that I propose that what is relevant is for personal information or sensorial space to remain *unaccessed*, as opposed to them being inaccessible. The adjective "unaccessed" is not found in any dictionary, but there is no suitable existing term to convey in one word the property of not having been accessed. "Inaccessible" denotes the property of not being able to be accessed, which is different from being accessible yet not actually accessed. Analogous differentiations exist in English, however, that use the same prefixes (e.g., indisputable/undisputed, inalterable/unaltered, etc.). An access theory that focuses on inaccessibility is bound to collapse into a control theory (because when something becomes accessible to others we lose control over it).

One might wonder what makes the two species—informational and sensory access—part of the genus of privacy. The unity of the category of privacy is founded on the notion

of being personally unaccessed and the kinds of interests we have in not being accessed in these ways by others. When others have personal access to us, we become vulnerable to them. In some cases, this vulnerability is invited, as when we accept losing privacy with regards to our romantic partner in exchange for intimacy. In other cases, privacy losses are uninvited because they give unwanted others power over us. For example, if an enemy knows where you live, or where you hurt, they can use that knowledge to harm you.

In contrast to mere privacy, the right to privacy is a matter of having control over our personal information and sensorial space. The right to privacy is concerned, not only with actual circumstances, but also with counterfactual ones, and not with the objects of privacy, but rather with the ways of getting at objects of privacy (e.g., spying as a way to get personal information about someone).

The good of privacy is a minimally demanding or actual one—it is one we either have in the here and now, or we do not. The right to privacy, on the other hand, is a right to a rich or robustly demanding good. Robustly demanding goods are ones that require counterfactual assurances. The right to privacy requires, not only that you not invade my privacy here and now, but that you would not invade my privacy in a range of relevant possible situations (e.g., if you stopped liking me, or if invading my privacy suddenly became profitable for you).

Rich goods have a structure that mirrors the Republican ideal of freedom. For Republicans, it is not enough for someone not to suffer actual interferences to be free. A slave might have a master that has never interfered with him and still not be free. As long as someone could interfere with one arbitrarily (i.e., with impunity), one is not free (Pettit 1996). Philip Pettit (2015) has used this structure to argue that goods such as love, virtue, and respect are also counterfactually demanding in this way. I wish to include the right to privacy in this list. What we call the right to privacy then, is technically a right to *robust* privacy, but I will keep on calling it the "right to privacy" for short.

The need to incorporate both access and control in a comprehensive theory of privacy has become all the more important in the digital age because much of the data that we lose may never be accessed by another human being. Intelligence agencies around the world and thousands of corporations might be collecting your personal data, storing it, and allowing algorithms to analyze it and make decisions about your life.

Access in the age of AI

There is a debate to be had about what counts as accessing data in the age of AI. Let's start with a paradigm example. In the case of a private diary, it is intuitive to think that reading and understanding the contents of the diary is what counts as access. The commonsense justification is that reading and understanding the diary is what leads someone to learn something new about another person, and that is the essence of the loss of privacy.

Now let's think about cases in which personal data gets stored in a digital format. If a person, say, an intelligence analyst, opens that data and "reads" it (e.g., if she reads emails, watches videos, or listens to audio recordings), that seems like a clear-cut case of accessing data. Similarly, if a prospective employer buys a file on someone from a data broker, and takes a look at sensitive information (e.g., that this person has a disease), that is a clear-cut case of accessing personal data. But what about cases in which no human being ever looks

at the raw data, but personal data is used to make decisions about people? For instance, what should we say about a case in which an AI sifts through personal data and decides on that basis to reject a loan application?

One possibility is to say that there is no loss of privacy in these cases because there is no moral agent who learns anything new about the data subject. We can no more lose privacy to a computer than we can lose privacy to an ant. However, given that some of the *effects* of algorithms sifting through data are similar to those of privacy losses (e.g., discrimination), we should treat AIs accessing personal data as if it were equivalent to a human being accessing that data.

Although the debate about what counts as access is important to establish when privacy has been lost, it is not as important for establishing when there has been a violation of the right to privacy. The mere act of storing personal data changes the balance of power between two parties. It also increases the likelihood of losses of privacy (that is, it increases the chances of someone having access to that data), which in turn leads us to explore why privacy is important.

Why Is Privacy Important?

Privacy is important because it shields citizens from harms that arise from exposure to others. These include (a) certain abuses of power that may come about as a result of other people having access to our personal life (e.g., discrimination, identity theft, security threats),[5] (b) the demands of sociality, (c) being judged and possibly ridiculed by others (and thus from self-conscious negative emotions such as shame and embarrassment), and (d) the discomfort of being watched, heard, and so on.

Having privacy allows us to cultivate different kinds of relationships. According to James Rachels, the value of privacy relies on the "connection between our ability to control who has access to us and to information about us, and our ability to create and maintain different sorts of social relationships with different people" (Rachels, 1975). You are probably a very different person with your students than you are with your spouse—and all parties are likely grateful for that. Having harmonious social lives would be impossible if we could know everything about everyone at all times, and if we acted in accordance with how we feel at every moment (Cocking & van den Hoven, 2018). Some degree of concealment, reticence, and non-acknowledgment is necessary to avoid unnecessary conflict in the public sphere (Nagel, 1998). Such limits protect both the individual (from undue judgment from other people) and the public sphere, which ends up being much less toxic if it only gets exposed to the more polished aspects of individuals, as opposed to the unadulterated versions.

Privacy is also important for political reasons. One of the greatest virtues of liberal democracy is its emphasis on equality and justice. No one is above the law, everyone has the same rights, everyone of age gets a vote, and everyone gets the opportunity to participate in democracy in more active ways—even the people who end up on the losing side of a vote. One of the greatest vices of the data economy is how it's undermining equality in various ways. The very essence of the personal data economy is that we are all treated differently, according to our data. It is because we are treated differently that algorithms end up being sexist and racist, for instance. It is because we are treated differently that we get

to see different content, which further amplifies our differences—a vicious cycle of otherness and inequality. No matter who you are, you should have the same access to information and opportunities. Personifications of justice are often depicted wearing a blindfold, symbolizing justice's impartiality. Privacy is what can blind the system to ensure that we are treated equally and impartially. Privacy is justice's blindfold.

For Whom Is Privacy Important?

Privacy is, first, important for individuals to protect themselves from abuses of power from governments (e.g., illegitimate surveillance), corporations (e.g., questionable profiling), and other individuals (e.g., public shaming). Furthermore, contrary to common belief, privacy is not only valuable individually, but also collectively. Privacy's ability to shield us from abuses of power makes it a common good, as abuses of power through violations of the right to privacy can jeopardize other common goods such as national security and the legitimacy of elections.

Individual harms

One set of harms that privacy protects us from is illustrated by revenge porn—the nonconsensual sharing of nude or sexual images—and related harms such as blackmail. Others' attention and judgment can cause people to feel self-conscious at best, and humiliated or shunned at worst.

Revenge porn is not uncommon. According to a survey of nearly 4,300 people, one in five Australians has been a victim of image-based abuse. In some cases, sensitive images get shared and exposed; in other cases, the threat of exposure is used to coerce, extort, or harass the victim (Henry et al., 2017).

Other individual harms include identity theft and fraud. A woman who acquired Ramona María Timaru's personal details used them to impersonate her and take out loans in banks across Spain that were never paid back. It is surprisingly difficult to prove you did not commit a crime when someone is committing them in your name. Timaru has been detained multiple times, and she has spent years and a substantial amount of money defending herself in many trials in different parts of Spain. When the newspaper *El País* interviewed her, she said that her life "was being ruined" and that she was taking tranquilizers to fight anxiety (Hernández, 2016).

Some other individual harms are more difficult to notice but can be just as damaging. One is discrimination. Data brokers are companies that strive to collect all the data they can on internet users. Information can include census and address records, driving records, web-browsing history, social media data, criminal records, academic records, credit records, medical records, and more. They then sell these files to banks, would-be employers, insurance companies, and governments, among others.

Imagine two candidates who are equally qualified for a particular job, but the data broker's file on one of them shows that he suffers from health issues. The company decides to hire the healthy candidate and tells the other one that there was someone more qualified for

the job. In theory, discrimination is illegal. In practice, it is very hard to prove; companies can always come up with untruthful explanations for their decisions, and victims may not even realize they have been discriminated against. Discrimination may take several forms: if your genetic information is not private, an insurance company that suspects you to have bad genes can charge more expensive premiums for something over which you have no control and for which you cannot be blamed.

Collective harms

Privacy damages can also be collective. In the 2018 Cambridge Analytica scandal, it was revealed that personal data from 87 million Facebook accounts had helped build psychological profiles of internet users who were then sent personalized political propaganda. Cambridge Analytica worked on both the 2016 U.S. election and the EU referendum campaign in Britain that same year. During the referendum, the firm was on the "leave" side; voters who were leaning towards voting "leave" got information that reinforced their views, including false news regarding immigration; voters who were thinking of voting "remain" might have been sent information that discouraged them from going to the ballot box. Propaganda is not new, but before the internet came along it used to be something public—everybody could see what each party was advertising. What is particularly unhealthy about personalized propaganda is that it contributes to polarization through showing each person different and potentially misleading information, and it takes advantage of people's personality traits to be more effective in influencing them. In the past, propaganda may have been just as misleading, but at least we all had access to it. Having journalists, academics, and ordinary citizens discussing the same material helped to put it into perspective. Personalized propaganda causes individuals to be blind to what others' are seeing. It fractures the public sphere into atomic individual spheres. The Trump campaign, for example, used six million different ads, targeted at different people.

One lesson of the Cambridge Analytica case is the collective nature of privacy. Privacy is not only collective because of the consequences of its loss—even if "only" 87 million Facebook users lost their privacy, all the citizens of the manipulated democracies were indirectly harmed. Privacy is also collective in another way: when you expose information about yourself, you inevitably expose others as well.

Only 270,000 Facebook users consented to Cambridge Analytica collecting their data. The other 87 million people were friends of the consenting users whose data was harvested without their knowledge. We are responsible for each other's privacy because we are connected in ways that make us vulnerable to each other. Think of all the contacts you have on your mobile phone. If you give a company access to that phone, you give it access to your contacts too. If you divulge genetic information, you expose your parents, siblings, and children. If you reveal your location data, you inform on people with whom you live and work. If you disclose your habits and psychological make-up, you expose people who resemble you.

The collective aspect of privacy has implications for shared undertakings, such as national security. A general loss of privacy can make it easy for rival countries to know too much about public officials, military personnel, and so on, which could facilitate an attempt to blackmail people, for instance (Thompson & Warzel, 2019).

Collective harms facilitated by privacy losses can be dramatic. The Nazis were more effective in finding Jews in countries in which civil registers had more detailed personal data (e.g., about religious affiliation and the addresses of family members). It is no coincidence that the Nazis were great innovators in techniques of registration and identification of individuals. Surveillance is always intimately associated with control. It is also worth noting—given the new power of technology companies—that it was a technology company, IBM, who assisted them in these objectives through the development of punch cards (Black, 2012).

What the collective nature of privacy suggests is that, in addition to there being a right to privacy, protecting one's own and others' privacy may be understood as a civic duty. Privacy is a public good that supports values like equality, fairness, and democracy, and it takes a collective effort to protect it.

Protecting Privacy in the Digital Age

In this section, I will tease out some of the practical implications of the hybrid account of privacy. I will end the section with some thoughts about consent.

The access account of privacy, in conjunction with the view that privacy is important to protect individuals and society from abuses of power, suggests that (other things being equal) we should do what we can to minimize losses of privacy—regardless of whether those losses amount to violations of the right to privacy. The control account of the right to privacy suggests that, other things being equal, the act of collecting and storing personal data can violate the right to privacy, even when the data is not accessed. Therefore, other things being equal, we should not collect or store personal without meaningful consent and without minimizing risks. From these conclusions we can infer a few practical implications.

In the current data landscape, some of the most effective ways of lessening losses of privacy would be through data minimization, storage limitation, and banning the trade in personal data.

Data minimization

Data minimization is the principle according to which one should only collect the minimum amount of personal data needed to fulfil some purpose. While this principle is part of the European General Data Protection Regulation (GDPR), it is rarely practiced seriously. That is partly because it is currently legal to sell personal data. Why would anyone minimize data collection when they can profit from it? No one likes minimizing one's profit. For data minimization to make sense, we have to ban the trade in personal data (more on that below) and stipulate what counts as a legitimate purpose for personal data collection. One possibility is to specify that personal data can only be collected for the purposes of benefitting data subjects (i.e., implementing fiduciary duties).

An effective way of ensuring data minimization is to set all default settings to "no data collection" or "minimal data collection," such that individuals would have to make a conscious effort to have their personal data collected. That would also have the much more

palatable result that people wouldn't need to constantly ask for their privacy to be respected (e.g., by saying "no" to tracking every single time they enter a website). And people who opt in to data collection could legitimately be remembered, so they would only need to do it once.

Storage limitation

A related principle to data minimization is that of storage limitation, which is also part of the GDPR. Storage limitation states that personal data must not be stored for longer than what is strictly necessary. Once again, if we allow the trade in personal data, it becomes "necessary," or rather, profitable, to store data indefinitely, as one can always hope to sell it. Therefore, here again, the trade in personal data must be banned, and we must specify what counts as acceptable purposes. Once acceptable purposes have been established, most personal data should not be collected without a plan to delete it.

Storing personal data indefinitely is a recipe for disaster for at least two reasons. First, personal data is sensitive, highly susceptible to misuse, hard to keep safe, and coveted by many—from criminals to insurance companies and intelligence agencies. The longer personal data is stored, the likelier it is that it will end up being misused. Second, forgetting plays an important social role.

Social forgetting provides second chances. Expunging old criminal records of minor or juvenile crimes, forgetting bankruptcies, and erasing the records of paid debts offer a second chance to people who have made mistakes. Societies that never forget tend never to forgive.

For most of history, keeping records has been difficult and expensive. Paper used to be extremely costly, and we needed a fair amount of space to store it. Writing demanded time and dedication. Such constraints forced us to choose what we wanted to remember. Only a tiny fraction of experience could be preserved, and even then, memory was shorter-lived than it is today. Back when paper was not acid-free, for instance, it disintegrated rather quickly. Such documents had a built-in expiry date set by the materials they were made of (Mayer-Schönberger, 2009).

The digital age has turned the economics of memory upside down. Today, it is easier and cheaper to remember it all than to forget. Once data collection became automated, and storage became so cheap that it was suddenly realistic to aspire to collect it all, we went from having to select what to remember to having to select what to forget. Because selecting takes effort, forgetting has become more expensive than remembering by default.

It is tempting to think that having more data will necessarily make us smarter, or able to make better decisions. In fact, it may impede our thinking and decision-making capabilities. Human forgetting is partly an active process of filtering what is important. Not selecting what we remember means that every piece of data is given the same weight, which makes it harder to identify what is relevant in a vast field of irrelevant data (Mayer-Schönberger, 2009).

We are collecting so much data that it is impossible for us to glean a clear picture from it—our minds have not evolved to process such massive amounts of information. When we have too much data and we're trying to make sense of it, we face two options. The first is to select a bit of information based on some criterion of our choosing that might make us blind to context in a way that can reduce our understanding, rather than increase it. The

second, and increasingly common, option to try to make sense out of inordinate amounts of data is to rely on algorithms as filters that can help us weave a narrative.

One challenge we face is that algorithms have no commonsense to know what is important in a sea of data. For instance, an algorithm designed to determine who is a criminal by analyzing facial images might end up picking out people who are not smiling. The algorithm doesn't have the necessary reasoning capacity to understand that, in its training data, the images of criminals provided by the police were ID photos in which people were not smiling.[6] Furthermore, algorithms have been shown time and again to suffer from biases embedded in our data, in the assumptions we make about what we are trying to measure, and in our programming.

Handling too much data, then, can lead to less knowledge and worse decision-making. The double risks of misinterpreting data in ways that obscure the truth and of memory impeding change combine to make permanent and extensive records about people dangerous. Such records capture people at their worst and don't allow them to transform into someone better. Old personal data can also trap us into historical biases: if we use old data to build the future, we will be prone to perpetuating past mistakes.

We need to introduce forgetting into the digital world. That is partly the spirit behind Europe's right to be forgotten. When Mario Costeja did a Google Search on his name in 2009, some of the first items to come up were a couple of notices from the late 1990s in the Spanish newspaper *La Vanguardia*. The notices were about Costeja's house being auctioned to recover his social security debts. They had first been published in the paper edition of the newspaper, which was later digitized.

Costeja went to the Spanish Data Protection Agency to complain against *La Vanguardia*. He argued those notices were no longer relevant because his debts had been settled. Having that stain linked to his name was hurting his personal and professional life. The newspaper had refused to delete the records, and the Spanish Data Protection Agency agreed with it—*La Vanguardia* had published those public records lawfully. But the agency did ask Google to delete the link to the auction notice. A person who has paid his debts should not be burdened with that weight for the rest of his life.

Google appealed the decision, and the case ended up in the European Court of Justice, which, in 2014, ruled in favor of the right to be forgotten. Costeja's records can still be found in *La Vanguardia* but they are no longer indexed in Google Search. Although the implementation of this right has given rise to doubts and criticism, its principle makes sense. A right to be forgotten protects us from being haunted by personal data that is "outdated, inaccurate, inadequate, irrelevant, or devoid of purpose, and when there is no public interest" (Powles & Chaparro, 2015).

Banning the trade in personal data

Even in the most capitalist of societies we agree that certain things are not for sale. We don't sell people, votes, organs, or the outcomes of sports matches. We should add personal data to that list. As long as we allow personal data to be sold, the incentive will be to amass it and store it indefinitely, and both of these policies maximize losses of privacy.

Today, personal data is collected by corporations for many different purposes (marketing and improving services among the top ones), and then it often gets sold on to data brokers.

A typical data broker will have thousands of data points about every person, including age, gender, education, employment, political views, relationship status, purchases, loans, net worth, vehicles owned, properties owned, banking and insurance policies details, likelihood of someone planning to have a baby, social media activity, alcohol and tobacco interests, casino gaming and lottery interests, religion, health status, and much more (Melendez & Pasternack, 2019). Data brokers go on to sell this data (through individual files or through categorized lists of people) to insurance companies, banks, prospective employers, and governments, among others.

Equifax is one of the largest data brokers and consumer credit reporting agencies in the world. Its data breach is one of the worst in corporate history (Hoffman, 2019). In September 2017, it announced a cybersecurity breach in which criminals accessed the personal data of approximately 147 million American citizens. The data accessed included names, social security numbers, birth dates, addresses, and driver's license numbers. It is one of the biggest data breaches in history. In February 2020, the story became even more troubling when the United States Department of Justice indicted four Chinese military people on nine charges related to the breach (which China has so far denied).

The very existence of sensitive files on internet users is a population-level risk. Many times, personal data held by data brokers is not even encrypted or well protected. Data brokers currently don't have enough of an incentive to invest in good security, which results in risks for society and individuals.

Buying profiles from data brokers is not expensive. Bank account numbers can be bought for 50 cents, and a full report on a person can cost as little as 95 cents (Dwoskin, 2014; Angwin, 2014). For less than $25 per month one can run background checks on everyone one knows. In May 2017, Tactical Tech and artist Joana Moll purchased a million online dating profiles from USDate, a dating data broker. The haul included almost five million photographs, usernames, email addresses, details on nationality, gender and sexual orientation, personality traits, and more. Although there is some doubt regarding the source of the data, there is evidence that suggests it came from some of the most popular and widely used dating platforms. It cost them €136 (about $150). That such a transaction is possible is astounding. Personal data is valuable, cheap, and sensitive—an explosive combination for privacy.

Part of what good regulation entails is stopping one kind of power turning into another. For instance, good regulation prevents economic power turning into political power (i.e., money buying votes, or politicians). In the same way, we need to stop the power accrued through personal data transforming into economic or political power. Personal data should benefit citizens—it shouldn't line the pockets of corporations at the expense of citizens or democracy.

Banning the trade in personal data does not mean banning the collection or proper use of such data. Some kinds of personal data are necessary (e.g., for medical treatment and research). But our health system should not be allowed to share that data, much less sell it.

Ending the trade in personal data does not mean that other kinds of data should not be shared—the ban need only apply to *personal* data. In fact, some non-personal data should be shared widely to promote collaboration and innovation. As computer scientist Nigel Shadbolt and economist Roger Hampson argue, the right combination is to have "open public data" and "secure private data" (Shadbolt & Hampson, 2019).

We need, however, stricter definitions of what counts as personal data. At the moment, legislation such as the GDPR does not apply to anonymized data. All too often, however, data that was thought to be anonymous has ended up being easily re-identified. Part of the problem is that we are not sure what techniques may be developed and used in the future to re-identify individuals in an "anonymous" database.

We also need to have a very broad understanding of what counts as a data trade. Data brokers provide personal data in exchange for money, but many other companies make data deals that are less crude. Facebook, for instance, has given other companies access to its users' personal data in exchange for these companies treating Facebook favorably on their platforms. Facebook gave Netflix and Spotify the ability to read its users' private messages, and it gave Amazon access to users' names and contact information through their friends. Part of what it received in return was data to feed its invasive friend-suggestion tool, "People You May Know" (Dance et al., 2018). Personal data should not be part of our commercial market. It should not be sold, disclosed, transferred, or shared in any way for the purposes of profit or commercial advantage.

The role of consent

In offline settings, we usually think that consented losses of privacy do not amount to violations of the right to privacy. If I willingly tell something private to my friend, no violation has occurred. The value of informed consent comes from the context of medicine, in which patients must give permission to doctors to receive treatment. Unfortunately, the power that we confer to consent in the offline world does not translate well to the online world of big data.

One of the most common defenses that big tech companies use against privacy criticisms is that users are consenting to the collection of their personal data. But the consent we give to data collection is typically neither freely given nor informed. It is also unclear whether individuals have the moral authority to consent to data collection. I'll analyze each of these difficulties in turn.

Consent is not freely given because often people do not feel like they have the freedom to opt out. Most tech companies have invasive terms and conditions, and opting out of them can often amount to opting out of being a full participant in one's society (e.g., getting an education, fulfilling one's duties at work, etc.).

Furthermore, consent to data collection is very rarely informed. Reading the entirety of the privacy policies of the websites you interact with would amount to a full-time job. Even if you had the time to read them, you would likely not understand the technicalities unless you are a data protection lawyer. Current consent practices put too heavy a burden on the shoulders of ordinary citizens. It's analogous to ask people to certify themselves that the food they buy in the supermarket is edible. Privacy policies are notorious for being documents designed, not to protect consumers, but to minimize the liability of companies. (Proof of this interpretation is that privacy policies often contain the disclaimer that the terms and conditions you are signing may change at any time.)

Even if you managed to understand the technicalities of what you are accepting, privacy policies are, more often than not, too vague to provide enough information for meaningful

consent (e.g., they allude to sharing data with unnamed "third parties"). Perhaps more important of all, not even data scientists could provide meaningful consent because big data is designed to reveal unforeseen correlations, which implies that there will be a significant degree of uncertainty about future findings. In other words, data subjects cannot be told about future uses and consequences of their data because not even researchers can know what kind of correlations may be unveiled, and they often cannot guarantee how this data will be used. This last problem could be somewhat ameliorated if data collectors establish an expiry date after which time the data will be deleted, and if they make explicit the kinds of inferences that they will be looking for.

A final problem with consent is that, given the collective aspect of privacy, it is not clear whether individuals have the moral authority to give away their personal data whenever that data contains personal about other people (e.g., as in the case of genetic data) or whenever the loss of privacy will have collective consequences (e.g., as in the case of Cambridge Analytica), all of which leads us to revisit the understanding that society has an interest in citizens protecting their privacy.

The right to privacy is a right of the individual against other people, corporations, and the state. Individuals have an interest in having privacy because it protects them from abuses of power. But society also has an interest in people protecting their privacy, because privacy protects collective values like equality, privacy, and democracy. In this sense, the right to privacy is similar to journalists' right to protect their sources. When journalists protect their sources, they protect a pillar of democracy.

The collective aspect of privacy is partly what makes it a fundamental right. In Joseph Raz's words, "fundamental moral rights cannot be conceived as essentially in competition with collective goods. On examination either they are found to be an element in the protection of certain collective goods, or their value is found to depend on the existence of certain collective goods" (Raz, 1988). The collective side of privacy is also one reason why personal data should not be thought of as private property—individuals do not have the moral authority to sell their data like they have a moral authority to sell their property (Véliz, 2020a).

The complexities of big data and AI are such that consent is a limited tool in protecting the right to privacy. Consent still has a role to play, but it cannot be what does most of the work in protecting citizens and society. We need other measures—like data minimization, storage limitation, and banning the trade in personal data—to better protect privacy. Where does that leave the notion of control in our account of the right to privacy? Control should be understood as negative control; that is, the capability of preventing someone who wants to gain access to our personal data from gaining that access (Mainz & Uhrenfeldt, 2021). Insofar as measures like data minimization limit the possibilities that someone might gain access to your data, they increase your control of that data. Other advancements in privacy can also empower consent in ways that go beyond what we have today. For example, currently, there is no easy way to withdraw consent from data collection. Once you have given consent, your data gets shared so widely, that by the time you try to withdraw consent, your data has been passed on and replicated multiple times. Something like Sir Tim Berners-Lee's project Solid—personal data pods in which data is stored and over which users have control—would allow people to instantly withdraw their data from any institution they have shared it with.

Conclusion

This chapter has explored the nature of privacy, its value, and ways to protect it. According to the hybrid account of privacy offered, privacy amounts to being personally unaccessed, and enjoying the right to privacy amounts to not losing (negative) control involuntarily over one's personal information or sensorial space. I argued that the value of privacy relies on it shielding citizens from harms that arise from exposure to others. Privacy, I argued, is not only important for individuals, but also for society. Finally, I proposed data minimization, storage limitation, and banning the trade in personal data as effective ways of minimizing losses of privacy and violations of the right to privacy in the digital age. The chapter ended by exploring the limits of consent in the context of AI and big data.

Notes

1. One notable account that doesn't seem to fit into either of these categories is Helen Nissenbaum's contextual privacy, according to which privacy is protected through respecting appropriate flows of information that conform with contextual information norms. One important problem with Nissenbaum's theory is that it relies too much on social norms. The theory does not give us any way to normatively assess whether current privacy norms are morally justified. Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
2. Some access-based theories of privacy include Allen, A. (1988). *Uneasy access: Privacy for women in a free society*. Rowman and Littlefield; Garrett, R. (1974). The nature of privacy. *Philosophy Today 18* , 263–284; Gavison, R. (1980). Privacy and the limits of law. *The Yale Law Journal 89* , 421–471; Gross, H. (1971). Privacy and Autonomy. In J. R. Pennock and J. W. Chapman (Eds.), *Privacy: Nomos XIII*. Atherton Press; Parent, W. A. (1983). Recent work on the concept of privacy. *American Philosophical Quarterly 20*.
3. I argue against self-presentation accounts of privacy in Véliz, C. (2022). Self-presentation and privacy online. *Journal of Practical Ethics 10*.
4. Marmor's view is one instance of control-based approaches to privacy. Other control theories include Fried, C. (1970). *An anatomy of values*. Harvard University Press; Bezanson, R. P. (1992). The right to privacy revisited: Privacy, news, and social change, 1890–1990. *California Law Review 80* , 1133–1175; Parker, R. (1974). A definition of privacy. *Rutgers Law Review 27*; Beardsley, E. (1971). Privacy: Autonomy and self-disclosure. In J. R. Pennock and J. W. Chapman (Eds.), *Privacy: Nomos XIII*. Atherton Press; Gerstein, R. (1978). Intimacy and privacy. *Ethics 89* , 86–91; Rachels, J. (1975). Why privacy is important. *Philosophy and Public Affairs 4* , 323–333; Reiman, J. (1976). Privacy, intimacy and personhood. *Philosophy and Public Affairs* 6, 26–44; and Wasserstrom, R. (1978). Privacy: Some arguments and assumptions. In R. Bronaugh (Ed.), *Philosophical law*. Greenwood Press.
5. For more on the relationship between privacy and power, see Véliz, C. (2020b). *Privacy is power*. Bantam Press.
6. I take this example from Carl Bergstrom and Jevin West's analysis of a paper that claims that an algorithm can determine whether someone is a criminal from analyzing a facial image: Criminal Machine Learning, https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html.

References

Allen, A. (1988). *Uneasy access: Privacy for women in a free society*. Rowman and Littlefield.

Angwin, J. (2014). *Dragnet nation*. Times Books.

Beardsley, E. (1971). Privacy: Autonomy and self-disclosure. In J. R. Pennock and J. W. Chapman (Eds.), *Privacy: Nomos XIII* (pp. 56–70). Atherton Press.

Bezanson, R. P. (1992). The right to privacy revisited: Privacy, news, and social change, 1890–1990. *California Law Review 80* , 1133–1175.

Black, E. (2012). *IBM and the Holocaust*. Dialog Press.

Cocking, D., & Van Den Hoven, J. (2018). *Evil online*. Wiley Blackwell.

Dance, G. J. X., Laforgia, M., & Confessore, N. (2018). As Facebook raised a privacy wall, it carved an opening for tech giants. *New York Times*.

Dwoskin, E. (2014). FTC: Data brokers can buy your bank account number for 50 cents. *Wall Street Journal*. https://www.wsj.com/articles/BL-DGB-39567

Fried, C. (1970). *An anatomy of values*. Harvard University Press.

Garrett, R. (1974). The nature of privacy. *Philosophy Today 89* , 421–472.

Gavison, R. (1980). Privacy and the limits of law. *The Yale Law Journal 89* , 421–471.

Gerstein, R. (1978). Intimacy and privacy. *Ethics 89* , 86–91.

Gross, H. (1971). Privacy and autonomy. In J. R. Pennock & J. W. Chapman (Eds.), *Privacy: Nomos XIII* (pp. 169–181). Atherton Press.

Henry, N., Powell, A., & Flynn, A. (2017). Not just "revenge pornography": Australians' experiences of image-based abuse. A Summary Report. RMIT University.

Hernández, J. A. (2016). Me han robado la identidad y estoy a base de lexatín; yo no soy una delincuente. *El País. https://elpais.com/politica/2016/08/23/actualidad/1471908298_138488.html*

Hoffman, D. A. (2019). Intel executive: Rein in data brokers. *New York Times. https://www.nytimes.com/2019/07/15/opinion/intel-data-brokers.html*

Mainz, J. T., & Uhrenfeldt, R. (2021). Too much info: Data surveillance and reasons to favor the control account of the right to privacy. *Res Publica 27* , 287–302.

Marmor, A. (2015). What is the right to privacy? *Philosophy and Public Affairs 43* , 3–26.

Mayer-Schönberger, V. (2009). *Delete*: *The virtue of forgetting in the digital age*. Princeton University Press.

Melendez, S. & Pasternack, A. (2019). Here are the data brokers quietly buying and selling your personal information. *Fast Company*. *https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information*

Nagel, T. (1998). Concealment and exposure. *Philosophy and Public Affairs 27* , 3–30.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Parent, W. A. (1983). Recent work on the concept of privacy. *American Philosophical Quarterly 20* , 341–354.

Parker, R. (1974). A definition of privacy. *Rutgers Law Review 27* (2), 275–297.

Pettit, P. (1996). Freedom as antipower. *Ethics 106* (3), 576–604.

Pettit, P. (2015). *The robust demands of the good*. Oxford University Press.

Powles, J., & Chaparro, E. (2015). How Google determined our right to be forgotten. *Guardian*, February 18.

Rachels, J. (1975). Why privacy is important. *Philosophy and Public Affairs 4* , 323–333.

Raz, J. (1988). *The morality of freedom*. Oxford University Press.

Reiman, J. (1976). Privacy, intimacy and personhood. *Philosophy and Public Affairs 6* , 26–44.
Shadbolt, N., & Hampson, R. (2019). *The digital ape: How to live (in peace) with smart machines*. Oxford University Press.
Thompson, S. A. & Warzel, C. (2019). Twelve million phones, one dataset, zero privacy. *New York Times*. *https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html*
Véliz, C. (2020a). *Data, privacy & the individual*. Center for the Governance of Change, IE University.
Véliz, C. (2020b). *Privacy is power*. Bantam Press.
Véliz, C. (2022). Self-presentation and privacy online. *Journal of Practical Ethics 10* , 30–43.
Wasserstrom, R. (1978). Privacy: Some arguments and assumptions. In R. Bronaugh (Ed.), *Philosophical law* (pp. 148–166). Greenwood Press.

CHAPTER 8

THE CONCEPT OF ACCOUNTABILITY IN AI ETHICS AND GOVERNANCE

THEODORE M. LECHTERMAN

Calls to hold artificial intelligence to account are intensifying. Activists and researchers alike warn of an "accountability gap" or even a "crisis of accountability" in AI.[1] Meanwhile, several prominent scholars maintain that accountability holds the key to governing AI.[2] Progress on accountability, they contend, will unlock solutions to numerous other challenges that AI poses, including algorithmic bias, the unintelligibility of algorithmic decisions, and the harmful consequences of certain AI applications.

Appeals to accountability among AI commentators reflect a general trend in public discourse that both lionizes the concept and struggles to specify what it means. Historians of accountability note that the term originates in practices of financial record-keeping and only entered mainstream usage in the late twentieth century.[3] It has since become an ever-expanding concept, used both for narrow purposes and as a catch-all term for normative desirability.[4] This conceptual fuzziness is on full display in AI debates, where no two commentators seem to use the term in precisely the same way. Some scholars treat accountability as a kind of master virtue, using accountability as more or less synonymous with moral justifiability.[5] According to this perspective, AI is accountable when all its features are justifiable to all concerned. Others assign accountability a far more limited role, such as to verify that algorithms comply with existing legal standards or that aspects of system performance are traceable.[6] Some understand accountability as a mechanism for regulating professional roles and organizational relationships;[7] others suggest that accountability is a basic component of moral responsibility that exists independently of institutional practices.[8] Some presume that accountability is a quality of those who design and deploy AI systems,[9] while others treat accountability as a quality of the systems themselves.[10] Because these conceptual disagreements are rarely made explicit, participants in debates about AI accountability often talk past each other.

This chapter begins by disambiguating some different senses and dimensions of accountability, distinguishing it from neighboring concepts, and identifying sources of confusion. It proceeds to explore the idea that AI operates within an accountability gap arising from technical features of AI as well as the social context in which it is deployed. The chapter also

evaluates various proposals for closing this gap. I conclude that the role of accountability in AI ethics and governance is vital but also more limited than some suggest. Accountability's primary job description is to verify compliance with substantive normative principles—once those principles are settled. Theories of accountability cannot ultimately tell us what substantive standards to account *for*, especially when norms are contested or still emerging. Nonetheless, formal mechanisms of accountability provide a way of diagnosing and discouraging egregious wrongdoing even in the absence of normative agreement. Providing accounts can also be an important first step toward the development of more comprehensive regulatory standards for AI.

Different Meanings of Accountability

Analyses tend to agree that accountability is a relational concept with multiple terms.[11] It involves some agent accounting *to* some other agent *for* some state of affairs *according to* some normative standard. A person can be accountable *to* their neighbors *for* noise pollution *according to* local ordinances and commonsense norms of decency. An employee can be accountable *to* an employer *for* the employee's work output *according to* the standards specified by their contract. But as these examples indicate, accountability can apply in wider and narrower senses, with different corresponding standards. A disrespectful neighbor transgresses general moral and legal standards, standards that apply regardless of any contractual agreement. Accountability can be understood, in this first instance, as a dimension of moral responsibility concerned with identifying the causes of states of affairs and assigning praise and blame. In such a case, we seek to the determine the source of the noise, assess whether wrongdoing has occurred, and apply any appropriate sanctions or demands for redress. By contrast, an employee's performance may have little to do with injured parties or independent standards of rightness or wrongness; holding the employee accountable involves the employer assessing the work against the terms of their contract. Accountability in this second instance is a more context-dependent quality. It arises within social practices and relationships where power is delegated from one party to another. Waldron refers to the first sense of accountability as *forensic accountability* and the latter as *agent accountability*.[12] I explore each of these senses of accountability in turn, while also registering the possibility of a third sense of accountability, *accountability-as-a-virtue*.

Accountability in the forensic sense is backward looking and relates closely to responsibility. Theories of responsibility seek to explain how individuals can be connected to their actions and the consequences of their actions in ways that make it appropriate to praise or blame them.[13] In common speech, responsibility and accountability are sometimes used interchangeably. But several philosophers understand accountability more specifically as a component of responsibility. On one prominent view, accountability refers to the conditions under which it is appropriate or fair to *hold someone responsible* for states of affairs.[14] Holding someone responsible is not the same as *believing that someone is responsible*. A victim of sustained psychological trauma may have their moral faculties blunted, leading them to commit a crime. We may believe this person to be responsible for the crime, in the sense that the act was theirs and they performed it with ill intentions. Still, we may believe the person not entirely accountable for the crime, because fully blaming or

sanctioning them would be unfair. To be accountable, according to this understanding, is to be susceptible to a demand for justification, to be expected to provide answers or render an account of what happened and why. It may also involve susceptibility to sanction if the justification comes up short.[15]

Importantly, this perspective holds that responsibility is a prerequisite for accountability: one cannot be accountable for a condition unless one is also responsible for that condition. Discussion of accountability and responsibility in the context of AI has tended to understand this relationship differently. Artificial agents can be accountable—i.e., can be susceptible to demands for justification or sanction—without necessarily being responsible or blameworthy. Floridi and Sanders argue that AI (at least in current and near-term forms) cannot be responsible for wrongdoing.[16] Much like nonhuman animals, AI cannot be responsible because it does not have the relevant intentional states. But AI can be accountable for wrongdoing because it can be sanctioned by modifying or deleting it. Similarly, in her classic study of accountability in a computerized society, Nissenbaum holds that responsibility is sufficient but not necessary to ground a demand for accountability.[17] An agent's being responsible for wrongdoing generates a reason to hold that agent accountable. But one can be accountable for a state of affairs without being responsible for it, such as if the state of affairs was caused by one's subordinate, one's pet animal, or one's technological artifact.

Discussions of accountability in the forensic sense treat accountability as a property to be attributed retrospectively in connection with discrete events. When a technological artifact is involved in some bad event, we seek to determine who or what is accountable for this, and to treat them accordingly. However, discussions of accountability as a dimension of responsibility also suggest that accountability might be understood as a virtue to be cultivated proactively.[18] An accountable individual, according to this understanding, is one who is robustly disposed to answer for their conduct, to welcome scrutiny of their decisions, and to take responsibility for harms. Likewise, an accountable agent or system is one that reliably welcomes input and oversight from relevant stakeholders, has the right features in place to ensure compliance with relevant standards, and fully acknowledges and rectifies its failures. *Accountability-as-a-virtue* is thus something one can display in greater or lesser quantities.

This way of conceiving accountability resonates with calls in popular discourse for AI—and those who design and deploy it—to be "more accountable." AI and those who deploy it often lack the qualities that enable interested parties to enjoy sufficient input, oversight, or redress. Accountability-as-a-virtue may also help to explain some of the conceptual confusion regarding accountability, as it represents a near antonym of forensic accountability. To be accountable in the forensic sense is generally a negative quality associated with blame and sanctions. But to be accountable in the virtuous sense is a positive quality associated with praise and rewards.

In addition to its usage in moral appraisal and legal investigation, accountability is often described as a more context-dependent quality tied to specific social practices.[19] Practices of accountability involve principal–agent relationships, where one party (the principal) delegates certain tasks or powers to another (the agent) and then monitors performance. The agent owes the principal accounts of this performance according to the terms specified by their relationship, which may be more or less explicit. This *agent accountability*, as Waldron terms it, is the dominant form of accountability within organizations, governments, and professional relationships.[20] It is also one way of characterizing the

relationship between citizens and public officials. Some go so far as to claim that this form of accountability is the "essence" of democracy, as it provides a way for those subjected to coercive power to constrain it.[21]

Problems arise when participants in an accountability relationship implicitly disagree about which model of accountability applies to a given situation. A helpful illustration is the accountability discourse around multilateral organizations. Grant and Keohane report that the World Bank is remarkably accountable to the governments that authorize it but remarkably unaccountable to those affected by its decisions.[22] According to the terminology suggested earlier, defenders of the Bank's accountability implicitly draw upon notions of agent accountability to assess the Bank's accounting to its principals, while critics draw on notions of forensic accountability to assess the Bank's treatment of other stakeholders. Because participants in these debates do not specify which sense of accountability they mean, productive deliberation stalls and tensions escalate.

As this example suggests, complex human societies can have various and overlapping practices of accountability, and confusion often arises over who has standing to demand an account, from whom, and for what. Thus, we can speak of the accountability of AI to its operators or creators, to its users or subjects, to lawmakers or regulators, and to society at large. We can speak of the accountability of AI designers and developers to their superiors, to the law, to industry standards, and to independent moral principles. AI may be perfectly accountable in one dimension but dramatically unaccountable in another dimension. It is not often clear which relationship applies or how subjects of accountability should prioritize amongst competing relationships.

Despite these challenges, the design and deployment of AI systems occur within a dense thicket of formal and informal accountability mechanisms, mechanisms that seek to facilitate the recording, reporting, evaluation, and sanctioning of decisions and activities. Generic accountability mechanisms in modern societies include legislation and law enforcement, the judiciary, government commissions, elections, auditors, whistleblowers, watchdogs, the press, certification standards, professional norms, compliance departments, and market forces, to name only a few. Each of these elements has a role in preventing transgressions of normative standards, diagnosing these transgressions, or sanctioning transgressions. As discussed below, accountability mechanisms also include a variety of tools proposed for AI specifically, such as transparency and explainability techniques, verification protocols, keeping humans in or on "the loop," and algorithmic impact assessments.

The AI Accountability Gap

The foregoing discussion suggests that two primary conditions need to be in place for accountability to be achieved. First, participants in an accountability relationship must have some basic agreement on the terms: who owes an account to whom for what and according to what standards. Second, subjects of accountability demands must be able to provide accounts according to these terms. Talk of accountability gaps, I propose, reflects a systemic problem with satisfying one or both conditions. A gap may emerge when participants disagree about whether they share an accountability relationship or what its terms are. A gap may also emerge when participants agree on the terms of the relationship but systematically

fail to uphold them for one reason or another. Several features of AI and its social context give rise to accountability gaps. These include (but are not limited to) the distribution of agency across humans and between humans and machines, the opacity and unintelligibility of algorithmic processes, and the persistence of moral and regulatory disagreement.

Distributed agency

One specific limitation to AI's accountability is the way this technology involves distributing agency across numerous human and nonhuman parties. Obviously enough, AI systems may involve the delegation of power from humans to machines. They also typically involve contributions from countless different parties, both human and nonhuman alike. I take up challenges with these features in turn.

The delegation of tasks by humans to autonomous machines involves relinquishing some degree of human control over outcomes. What makes AI novel and valuable is that it provides ways of thinking and acting without human direction and in ways that may be unforeseen by humans. An autonomous vehicle may take us to our destination on a route we never expected; an autonomous weapons system may identify a threat that its human colleagues never considered; AI may diagnose diseases, identify celestial objects, and predict weather all more accurately and more quickly than humans relying on traditional methods. These features are welcome when AI operates in ways that are consistent with human aims and interests, to optimize resource distribution, unravel scientific mysteries, and automate laborious tasks. But precisely because AI is to some extent independent from human understanding and control, it risks acting in ways that are unaccountable to its designers or operators and inconsistent with human aims and interests.

Recent criticism of AI's biased treatment of decision subjects or harms to society reveal that AI's accountability obligations are not limited to its designers and operators. Those who suffer adverse treatment from AI are entitled to demand an account and seek redress. The absence of avenues for appeal and redress of adverse algorithmic decisions is a glaring source of injustice. However, as some observers note, it remains pivotally important that AI be accountable to its designers and operators in the first instance.[23] If AI is not accountable to its designers and operators, it cannot be realistically accountable to anyone else.

This risk of an accountability gap between AI and its human overseers becomes graver the more advanced AI becomes, as there is the potential for AI to reach a level of sophistication where it begins to prioritize its own survival at the expense of human interests.[24] These prospects are somewhat remote, but they could very well be catastrophic. The risk of autonomous action also becomes grave when AI is used for high-stakes and irreversible applications, such as the exercise of lethal force. Faced with an opportunity to win a war, AI-directed weapons might raze an enemy's cities or eviscerate their own side's human soldiers caught in the crossfire.[25]

AI accountability also faces the problem of "many hands," the notion that AI decisions are ultimately the product of a vast number of different contributions from human and nonhuman agents alike.[26] Algorithmic systems often draw upon third-party datasets and a variety of third-party software elements, both proprietary and open source. The provenance and qualities of these elements may be unknown. Numerous individuals contribute to the collection and classification of data. Numerous further individuals contribute to the design,

testing, and deployment of models, which can be recombined and repackaged over time. The introduction of autonomous operations at various points throughout this sequence further obscures lines of attributability. In some cases, the problem of many hands is a problem because participants failed to record their specific contributions, and the problem could be reduced by requiring better record-keeping. In other cases, the number of operations in a causal chain may be so extensive or so convoluted that it is practically impossible to disentangle individual contributions to final outcomes. Of course, the problem of many hands is not specific to AI. Virtually every product in a modern economy arises from a complex chain of events and contributions before it is consumed. But AI-based products may foster this problem to a greater extent than others, owing to their particular complexity and the fact that some of these hands are not human.

Opacity and unintelligibility

Further accountability risks come from the facts that AI processes are often opaque to human observers, and even when they are more transparent, their decisions are often unintelligible to humans. AI systems may be based on faulty or biased data, they may contain errors in code, and they made encode controversial judgments of their designers. But those who interact with AI systems may not fully understand their purposes, how they work, or what factors they consider when making individual decisions. The problems arise not only from the scale and complexity of AI systems, but also from the proprietary nature of many components. In 2016, a civil society organization exposed that U.S. judges were using a biased algorithmic tool for making sentencing decisions, a tool whose criteria they did not understand—at least partly because the methodology was proprietary.[27]

A related problem arises in the interpretability of algorithmic decisions by those affected by them. Even when the source code and underlying data are available for scrutiny, the rationales for decisions may be difficult to interpret for experts and laypersons alike. The subject of an adverse decision by an algorithm may have little basis for assessing whether the process treated their case appropriately. The unintelligibility of algorithmic decisions inhibits the giving and receiving of accounts.

Moral and regulatory disagreement

Other things equal, accountability is more likely to be achieved in situations where there is already widespread agreement about normative standards—about what agents are accountable *for*. Consider some contrasts. Commonsense morality provides us with a basic shared understanding of the norms of friendship, which in turn allows us to hold friends to account when they fail to uphold these norms. Tort law and environmental regulations provide specific standards for negligence and pollution levels. Victims of a chemical plant disaster may hold the parties responsible for this disaster to account. However, especially when it comes to emerging technologies like AI, standards of harm and wrongdoing are often immature, unclear, or controversial. People disagree profoundly about the general ethical principles with which AI should comply. For instance, is it permissible for AI to reproduce inequalities in underlying conditions but not intensify them? Or should it be required to counteract these

background inequalities in some way? Should AI seek to nudge users toward complying with certain ideals of wellbeing, or should it err on respecting the liberty of its subjects?

Disagreements about the ethics of AI build upon more general disagreement about the nature of specific values like liberty and equality. They also build upon longstanding disagreements in normative ethics concerning how to appraise rightness or wrongness in general. When we hold AI accountable, should we take primary concern with intentions, actions, or results?[28] A credit-rating algorithm might be designed with the beneficent intentions of expanding access to credit, reducing capricious judgments by loan officers, and reducing loan default rates. Once deployed, it may in fact achieve these results. Despite these good intentions and results, it may also treat certain people unfairly. Different theories and different people place different weight on the significance of intentions, actions, and results in judging rightness and wrongness. When there is widespread disagreement about the standards to account *for*, generic calls for greater accountability appear to lack a clear target. In such cases, I suggest, calls for accountability might be understood as prompts for clarifying the normative standards that are prerequisites for successful accountability practices.

In addition to disagreement about which moral standards apply to AI, there is also tremendous disagreement over which regulatory standards apply to AI. Laws and industry conventions are still embryonic and competing for dominance. National governments and intergovernmental organizations have proposed many regulatory frameworks but so far passed little legislation. Seemingly, every professional association, standard-setting organization, and advocacy group is hawking a different list of principles, guidelines, and tools for regulating AI.[29] These efforts indicate broad agreement on the significance of the ethical challenges that AI poses. And many of these efforts reflect similar themes. But unless or until there is consolidation of competing terms, principles, and protocols, the accountability of AI is likely to suffer. Paradoxically, an abundance of competing standards can reduce accountability overall by inviting confusion and creating opportunities for actors to pick and choose the standards that burden them least.[30]

Closing the AI Accountability Gap

There are many proposals for closing aspects of AI's accountability gap. Some are more promising than others. Less promising proposals include attempts to ban broad categories of AI, initiatives to regulate AI as a general class, demands to make AI transparent, and proposals to make technology professionals the primary guardians of AI ethics. Alternatives include contextually sensitive regulatory approaches that appreciate differences in technological functions and domains of application; traceability and verification techniques; and a division of labor that expects professionals to flag ethical dilemmas without having the exclusive authority for adjudicating them.

Moratoria

One way to close an AI accountability gap is to eliminate its very possibility. Some suggest banning AI altogether or banning AI in entire domains of application, such as defense,

healthcare, or transportation.[31] Certain governments, including those of San Francisco, Oakland, Seattle, and Morocco, have enacted temporary bans on facial recognition until appropriate regulations can be devised.[32] Although narrowly crafted bans may indeed be warranted in cases like these, categorical bans face particular objections. One is that they may exceed their justifiable scope. Certain uses of AI in military applications or healthcare may be far less risky, or far more susceptible to accountability, than others. Automating target selection and choice of means is one thing. Using AI to assist human decision-makers about these things is another. And using AI for non-combat purposes, such as to optimize logistics or triage in humanitarian crises, is another thing altogether. Similarly, automating medical diagnoses and treatment decisions for life-threatening conditions may indeed create unacceptable risks. But these risks may not arise to the same extent when using AI to assist doctors in low-stakes diagnostic questions.[33]

Proposals to ban technologies must also be sensitive to the possibility of prisoners' dilemmas that may counteract a ban's intended effects. If Country A bans AI for military use, Country B gains a strategic advantage by continuing to develop AI for military use. If both countries agree to ban AI for military use, fear that either one may covertly continue development provides an incentive for each to continue development in secret. And even if effective monitoring mechanisms can make these commitments credible, there is always the possibility of a black market of non-state actors developing killer robots for the highest bidder.[34] If autonomous weapons are likely to be developed no matter what Country A does, banning all research and development may be self-defeating as it would put Country A at greater risk from attacks from other countries' autonomous weapons.

A third problem is that bans necessarily involve foregoing the potential benefits of AI, which can be tremendous. Automobile accidents kill 36,000 people in the United States each year, driven significantly by speeding, distraction, and driving under the influence.[35] Autonomous vehicles, which do not suffer from these problems, are expected to dramatically reduce road deaths, even as they may introduce or exacerbate other problems. Regulatory discussions that endorse a "precautionary principle" can fail to appreciate the opportunity costs of preserving the status quo.[36] Victims of traditional car crashes who would otherwise survive should autonomous vehicles be introduced have a powerful objection to postponing or preventing the deployment of self-driving cars.

Ironically, certain problems that moratoria aim to fix might be reduced by allowing a system to iterate and dynamically improve in large-scale applications. Thus, a diagnostic AI trained on biased data becomes dramatically more accurate the more patient data it receives. Sometimes this data can only be made available by releasing the product publicly. One difficulty here, of course, is the fairness to first-generation users of new technology, who must bear the consequences of less reliable products. But this problem is hardly unique to AI, and solutions to it have a long history in public health, for instance.

Some, but certainly not all, of the concerns animating calls to ban applications of AI stem from fully automated uses in which humans are not directly involved in the decision-making process or absent entirely. The now-familiar typology distinguishes between having humans "in the loop" (receiving advice from AI but responsible for determining whether and how to act on that advice), "on the loop" (where AI implements its own decisions while humans monitor and intervene if necessary), and "out of the loop" (where humans are not actively involved in deciding or monitoring).[37] In some cases, the accountability gap shrinks by keeping humans more closely involved and using AI primarily to augment human

intelligence rather than replace it completely.[38] As discussed further below, we have reason to worry about whether the humans in the loop are themselves the appropriate decision-makers, as those who design or operate AI and those who suffer the consequences of AI decisions are often not identical. But this issue is in principle separate from the question of human control itself.

Regulatory approaches: All-purpose and contextual

The steady stream of alarming mistakes and doomsday scenarios reveal the limits of patchwork regulatory standards and prompt increasing calls to regulate AI as a general class. Tutt, for instance, proposes a new federal agency modeled after the U.S. Food and Drug Administration to oversee the testing and approval of algorithms.[39] Such proposals would require imposing a common set of normative standards, technical criteria, and/or reporting requirements on all forms of AI. This would certainly make AI more formally accountable, but it would come with significant tradeoffs. AI is not monolithic and varies tremendously in its moral risks. Calls to regulate AI as a general class can fail to appreciate that many uses of AI are largely privately regarding and contain limited risks of harm. Consider AI applications for composing music.[40] Such technology might introduce or intensify disputes over intellectual property, but it raises no obvious threats to health, safety, or equality, and the case for granting the state additional oversight here appears relatively weak. Undifferentiated demands for public accountability can infringe on behavior that is more or less benign and privately concerned.

Although AI for music composition and AI for judicial sentencing clearly occupy opposite poles on the private-public scale, many applications of AI occupy a more nebulous intermediate area. Recommender algorithms on search engines and social media platforms are cases in point. Search engines and social media platforms are private corporations, but they can come to monopolize the flow of information with dramatic effects on public discourse and political stability.

A more promising approach to regulation would take account of various contextual factors, such as the domain of operation, the kinds of agents involved, asymmetries in information and power, and the different interests at stake. Different standards might apply based on whether subjection to the decisions is voluntary or nonvoluntary, whether the decisions are high-stakes or low-stakes, whether the risks of externalities are high or low, the degree of human oversight, the degree of competition, and so on. This idea has much in common with Nissenbaum's noted theory of privacy as "contextual integrity," a view holding that privacy is not an independent value but one that demands different things in different settings.[41]

Transparency and explainability

Talk of closing the accountability gap often appeals to principles of transparency and explainability.[42] Improving the transparency and explainability of AI is often claimed to be a major component of improving accountability, as we seem unable to determine whether AI complies with the reasons that apply to it if we cannot understand what it decides and

why. There is certainly a role for improvements in both qualities in making AI more accountable. But singular focus on either element leads to certain traps. As Kroll has argued, transparency is often neither desirable nor sufficient for making AI accountable.[43] It is not desirable in uses that require the protection of data subject privacy, trade secrets, or national security. It is not sufficient in most cases, as merely being able to view the data set or code of an algorithm is hardly a guarantee of making sense of it. When Reddit released code to the public indicating how its content moderation algorithm works, prominent computer scientists could not agree on how to interpret it.[44]

Demands for transparency often appear rooted in the implicit belief that transparency conduces to explainability or interpretability. If we can view the data or the code, this thinking goes, we are more likely to understand the algorithm's decisions. Although there continues to be interesting research and experimentation on improving the explainability of algorithmic decisions, to a certain extent the search for explainability is chimerical. The most advanced forms of AI are not programmed by humans but rather result from deep learning processes, which automatically create and adjust innumerable settings in response to training data. What these settings mean and why they were selected may be virtually unknowable. The more complex AI becomes, the harder its processes are to understand, and efforts to reverse-engineer them come with their own biases and limitations.[45] In situations where precise explanation is essential to the justification of a decision, as in criminal sentencing, it may be wiser to regulate the use of AI than to demand explainability from AI. Indeed, some propose that in high-stakes or public administration settings, the use of "black box" AI models is simply impermissible.[46] Decision-makers in these settings may only permissibly rely upon algorithmic tools that are interpretable by design and sufficiently well understood by their designers and operators.

A variety of alternative methods have been proposed to monitor the integrity and reliability of algorithmic systems in the absence of transparency and explainability. These include the banal but often overlooked methods of robust documentation and record-keeping,[47] clear divisions of responsibility during the development process, publicizing and following a set of standard operating procedures,[48] and different visualization and reporting methods to track and communicate the qualities of an algorithm, such as dashboards and "nutrition labels." Of particular interest to many are algorithmic impact assessments, which seek to forecast, declare, and offer mitigation strategies for potential adverse effects of a given AI application.[49] More technical tools include software verification techniques that check whether software matches it specifications and the use of cryptography to authenticate features and performance.[50] These methods cannot make AI fully explainable, but they can provide grounds for greater confidence in the results of AI decisions in certain cases.

Duty of care

Another proposed solution to the AI accountability gap involves tasking those who design and deploy AI with a duty of care to mitigate ethical risks of AI systems.[51] Many risks of AI can indeed be mitigated by heightened sensitivity of designers and operators to ethical issues. Greater awareness of structural injustice and the kinds of biases that may lurk in training data might be enough to prevent certain horrendous mistakes like the release of facial recognition products that classify Black faces as gorillas.[52]

However, a duty of care can be easily abused. Many ethical issues are too complex to be solved without more advanced expertise, and the ethical hubris of many technology professionals is already legendary.[53] Inviting professionals to take responsibility for ethically safeguarding their products can be a recipe for well-meaning mistakes, motivated reasoning, or encoding parochial value judgments into software. Many ethical issues arguably exceed the authority of technology professionals to resolve on their own. Plenty of these demand input from affected communities and a fair process of public deliberation. Overzealous exercise of the duty of care may invite criticism of paternalism or technocracy.

Some suggest that objections to the private governance of AI can be mitigated by limiting the range of eligible justifications for AI designs and outcomes. The criteria we use to appraise AI must operate within the bounds of "public reason"—reasons that any and every citizen could be expected to endorse.[54] This solution may certainly help to screen out the most parochial or controversial justifications, such as those rooted in narrow conceptions of human flourishing or faulty logic. But much, if not most, of the current disagreement in AI ethics already operates within the realm of public reason and appeals to public reason are of little help in resolving these debates.

An alternative approach to a duty of care is to train technology professionals on identifying and flagging ethical issues to be adjudicated by others. Designers and operators are the first line of defense in detecting potential harms from AI. With training, they may become attuned to noting the presence of controversial assumptions, disparate impacts, and value trade-offs. But deeper sensitivity to these ethical risks and appropriate ways of resolving them may profit from interdisciplinary collaboration between computing professionals and experts from academia and civil society. It is also a ripe opportunity for experimentation with new forms of civic engagement that allow input on technical questions by those affected by them.[55]

AI as an Accountability Instrument

The foregoing discussion has explored some of the challenges of ensuring that AI and those who design and apply it are accountable. However, it also pays to consider how AI might both erode and improve the accountability of conventional entities. AI can enable malicious actors and systems to evade accountability. It can also serve as an instrument for facilitating the accountability of humans and institutions.

Although not an instance of AI, blockchain is an adjacent form of digital technology that exemplifies this duality. A blockchain is a distributed ledger that uses cryptography to store value, facilitate exchanges, and verify transactions. Blockchain has applications in the verification of identities, the storage of digital assets, the assurance of contract fulfillment, and the security of voting systems. It creates strong mutual accountability by reducing reliance on individual trust or third-party institutions like governments, lawyers, and banks. Blockchain is most well-known for its use in cryptocurrency, decentralized media of exchange that are not authorized or controlled by central banks. Cryptocurrency is especially helpful to people and places ill-served by fiat currencies, where banking services may be inaccessible, dysfunctional, or discriminatory. But cryptocurrency's ungovernability creates a double-edged sword. An ungovernable currency becomes the medium of choice for illicit

transactions.[56] Furthermore, while major players can influence elements of the market, ordinary cryptocurrency holders have no way of holding the system to account for adverse conditions.[57]

The ways that AI can facilitate state surveillance and law enforcement are increasingly apparent in the forms of predictive policing, facial recognition, and judicial sentencing algorithms. Naturally, AI has numerous beneficial applications in government and can promote decisions that are more just and legitimate. In theory, decisions driven by rigorous data analysis can result in outcomes that are more efficient, consistent, fair, and accurate. Given optimistic assumptions about its ability to overcome challenges of bias and opacity, AI may even improve government accountability by reducing reliance on human discretion. But by expanding the power of states for surveilling subjects, controlling populations, and quashing dissent, AI also supplies states with powerful means for evading accountability.

As Danaher discusses (albeit skeptically), AI can also be part of the solution to state oppression by powering "sousveillance" methods that hold powerful actors to account.[58] Sousveillance refers to watching from below, and it is exemplified by efforts to film police misconduct on smartphones. AI in the hands of citizens and civil society groups may facilitate sousveillance by enabling the powerless to analyze data for signs of misconduct. This might take the form of journalists pursuing freedom of information requests, criminal justice advocates analyzing forensic evidence for signs of false convictions, or human rights activists tracking abuse through posts on social media. The tools of sousveillance also extend to consumer protection, as with applications that use bots to challenge bank fees, product malfunctions, and price gouging.

Conclusion: Accountability's Job Specification

We should be wary of placing too much faith in accountability as such. Greater accountability does not necessarily lead to greater justice. Functionaries who faithfully comply with the dictates of a genocidal regime are eminently accountable, in many respects. Software engineers might be perfectly accountable to their superiors, whose aim is to maximize profits at any social cost.

Some suggest that accountability is the "essence" of democracy, as it provides a constraint on unchecked power.[59] This position, however, also finds support with certain skeptics of democracy, who have sought to limit participation in politics to periodic opportunities to check abuses of power without opportunities to exercise or influence power in the first place.[60] For proponents of a more demanding view of the democratic ideal, accountability is better understood as but one feature of democratic legitimacy: namely, a condition on policy outcomes. For these perspectives, democratic legitimacy also requires conditions on policy inputs, such as collective self-determination, political equality, and deliberative decision-making.[61]

Debating accountability and its mechanisms can also distract us from fundamental questions about substantive normative standards. If we do not adequately address the question of what principles should regulate the design and use of AI and under what

conditions, debate about whether and how AI can be accountable to those principles seems to lose much of its point.

Despite accountability's limitations, however, the claim that accountability holds the key of AI ethics and governance is worth taking seriously. Especially when there is disagreement about substantive standards, accountability mechanisms may play an essential role in the discovery of problems and the search for more lasting solutions.[62] Procedural regularity, documentation, and impact assessments enable accounts to be given. There may be disagreement about the normative standards that apply to these accounts, but having the accounts is a critical step toward diagnosing problems, refining standards, and sanctioning failures. While accountability may not be all that democracy demands, institutional, organizational, and technical mechanisms that enable scrutiny of power are absolutely crucial to the protection and realization of democratic ideals.

Moreover, agreement on the details of principles of justice is not necessary for seeking accountability for violations of basic human rights and other obvious harms. There is already widespread agreement about certain fundamental rights and duties, and not all grounds for disagreement are reasonable. Improving the accountability of AI to basic moral standards would leave much work to be done, but it would also constitute a remarkable achievement.

Still, as this chapter has emphasized, the concept of accountability contains many puzzles and remains poorly understood. Improving the accountability of AI may be difficult to achieve without further work to disentangle and narrow disagreement on the concept's different meanings and uses.

Acknowledgments

For extraordinarily helpful comments on earlier drafts, I thank Johannes Himmelreich, Juri Viehoff, Jon Herington, Kate Vredenburgh, Carles Boix, David Danks, audience members at the 2021 Society for the Philosophy of Technology annual conference, and four anonymous readers for Oxford University Press.

Notes

1. Institute for the Future of Work. (2020). Mind the gap: How to fill the equality and AI accountability gap in an automated world. London, October; Raji, Inioluwa Deborah et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain, pp. 33–44, https://doi.org/10.1145/3351095.3372873; Hutchinson, Ben et al. (2020). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *ArXiv:2010.13561 [Cs]*, October 22, http://arxiv.org/abs/2010.13561.
2. Wachter, Sandra, Mittelstadt, Brent, & Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics 2* (6), eaan6080, https://doi.org/10.1126/scirobotics.aan6080; Skelton, Sebastian Klovig. (2019). Accountability is the key to ethical artificial intelligence, experts say. *ComputerWeekly.Com*, December 16, https://www.computerweekly.com/feature/Accountability-is-the-key-to-ethical-artificial-intelligence-experts-say.

3. Bovens, Mark et al. (2014). Public accountability. In Mark Bovens, Robert E. Goodin, & Thomas Schillemans (Eds.), *The Oxford handbook of public accountability* (pp. 1–20). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199641253.013.0012.
4. Mulgan, Richard. (2000). "Accountability": An ever-expanding concept? *Public Administration 78* (3), 555–573, https://doi.org/10.1111/1467-9299.00218.
5. See, e.g., Dignum, Virginia. (2020). Responsibility and artificial intelligence. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 218). Oxford University Press.
6. Kroll, Joshua A. et al. (2016). Accountable algorithms. *University of Pennsylvania Law Review 165* , 633; Kohli, Nitin, Barreto, Renata, & Kroll, Joshua A. (2018). Translation tutorial: A shared lexicon for research and practice in human-centered software systems. 1st Conference on Fairness, Accountability, and Transparency, New York.
7. Bovens, Mark. (2007). Public accountability. In Ewan Ferlie, Laurence E. Lynn, Jr., & Christopher Pollitt (Eds.), *The Oxford handbook of public administration* (pp. 182–208). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199226443.003.0009.
8. See, e.g., Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics 24* (2), 227–248, https://doi.org/10.5840/philtopics199624222; Shoemaker, David. (2011). Attributability, answerability, and accountability: toward a wider theory of moral responsibility. *Ethics 121* (3), 602–632, https://doi.org/10.1086/659003.
9. Nissenbaum, Helen. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2* (1), 25–42, https://doi.org/10.1007/BF02639315.
10. Floridi, Luciano, & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines 14* (3), 349–379, https://doi.org/10.1023/B:MIND.0000035461.63578.9d.
11. Goodin, Robert E. (2003). Democratic accountability: the distinctiveness of the third sector. *European Journal of Sociology 44* (3): 359–396, https://doi.org/10.1017/S0003975603001322; Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics 24* (2), 227–248, https://doi.org/10.5840/philtopics199624222.
12. Waldron, Jeremy. (2016). Accountability and insolence. In *Political political theory: Essays on institutions* (pp. 167–194). Harvard University Press.
13. Noorman, Merel. (2018). Computing and moral responsibility. In Edward N. Zalta (Ed.), *Stanford encyclopedia of philosophy*, https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.
14. Watson, "Two faces of responsibility."
15. Shoemaker, David. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics 121* (3), 602–632, https://doi.org/10.1086/659003.
16. Floridi & Sanders, "On the morality of artificial agents."
17. Nissenbaum, "Accountability in a computerized society."
18. Bovens, Mark. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics 33* (5), 946–967, https://doi.org/10.1080/01402382.2010.486119.
19. Goodin, "Democratic accountability"; Bovens, "Public accountability"; Bovens et al., "Public accountability"; Waldron, "Accountability and insolence."
20. Waldron, "Accountability and insolence."
21. Bovens, "Public accountability."
22. Grant, Ruth W., & Keohane, Robert O. (2005). Accountability and abuses of power in world politics. *American Political Science Review 99* (1), 29–43, https://doi.org/10.1017/S0003055405051476.

23. Wagner, Ben. (2020). Algorithmic accountability: Towards accountable systems. In Giancarlo Frosio (Ed.), *The Oxford handbook of online intermediary liability* (pp. 678–688). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780198837138.013.35.
24. Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.
25. Asaro, Peter. (2020). Autonomous weapons and the ethics of artificial intelligence. In S. Matthew Liao (Ed.), *Ethics of artificial intelligence* (pp. 212–236). Oxford University Press, https://doi.org/10.1093/oso/9780190905033.003.0008.
26. Nissenbaum, "Accountability in a computerized society."
27. Noorman, "Computing and moral responsibility."
28. Goodin, "Democratic accountability."
29. In 2019, one study counted 84 different initiatives to articulate ethical principles for AI. See Mittelstadt, Brent (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1* (11), 501–507. As of July 2021, a repository at www.aiethicist.org contained hundreds of different governmental and nongovernmental proposals for defining and upholding AI norms.
30. For a study of how the existence of overlapping accountability demands can go awry, see Koppell, Jonathan G. S. (2005). Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public Administration Review 65* (1), 94–108, https://doi.org/10.1111/j.1540-6210.2005.00434.x.
31. Asaro, "Autonomous weapons and the ethics of artificial intelligence."
32. Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). Algorithmic accountability for the public sector, August, https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/, p. 16.
33. Of course, the line between a low-stakes and high-stakes diagnostic question in medicine is often fuzzy, as symptoms of serious illness may often present similarly to symptoms of superficial illness.
34. The concern here is that unlike nuclear weapons and other weapons of mass destruction, the development of autonomous weapons has low barriers to entry and may be far more difficult to monitor. See, e.g., McGinnis, John O. (2010). Accelerating AI. *Northwestern University Law Review 104* (3), 1253–1269.
35. Insurance Institute for Highway Safety/Highway Loss Data Institute. (2021, March). Fatality Facts 2019: Yearly Snapshot, https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot.
36. Sunstein, Cass R. (2003). Beyond the precautionary principle. *University of Pennsylvania Law Review 151* (3), 1003–1058, https://doi.org/10.2307/3312884.
37. For an overview, see Rahwan, Iyad. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology 20* (1), 5–14, https://doi.org/10.1007/s10676-017-9430-8.
38. This is not to say that keeping humans in the loop is a panacea. The tendency of humans to trust too readily in the judgments of machines is a well-known source of cognitive bias. See, e.g., Skitka, Linda J., Mosier, Kathleen, & Burdick, Mark D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies 52* (4), 701–717, https://doi.org/10.1006/ijhc.1999.0349.
39. Tutt, Andrew. (2017). An FDA for algorithms. *Administrative Law Review 69* (1), 83–123.

40. See, e.g., Fernandez, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research 48* , 513–582, https://doi.org/10.1613/jair.3908.
41. Nissenbaum, Helen. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
42. Doshi-Velez, Finale et al. (2019). Accountability of AI under the law: The role of explanation. Preprint. *ArXiv:1711.01134 [Cs.AI]*, December 20, http://arxiv.org/abs/1711.01134.
43. Kroll, Joshua A. (2020). Accountability in computer systems. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (179–196). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780190067397.013.10.
44. New, Joshua, & Castro, Daniel. (2018). How policymakers can foster algorithmic accountability. Center for Data Innovation, May 21.
45. Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1* (5), 206–215, https://doi.org/10.1038/s42256-019-0048-x.
46. Ibid. See also, Busuioc, Madalina. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review 81* (5), 834.
47. Raji et al., "Closing the AI accountability gap."
48. Kroll et al., "Accountable algorithms."
49. For critical discussion, see Selbst, Andrew D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology 35* (1), 117–191.
50. Ibid.
51. Nissenbaum, "Accountability in a computerized society."
52. Guynn, Jessica. (2015). Google photos labeled Black people "gorillas". *USA Today*, July 1.
53. Morozov, Evgeny. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.
54. Binns, Reuben. (2018). Algorithmic accountability and public reason. *Philosophy & Technology 31* (4), 543–556, https://doi.org/10.1007/s13347-017-0263-5.
55. Landemore, Hélène. (2020). *Open democracy: Reinventing popular rule for the twenty-first century*. Princeton University Press.
56. Foley, Sean, Karlsen, Jonathan R., & Putniņš, Tālis J. (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies 32* (5), 1798–1853, https://doi.org/10.1093/rfs/hhz015; Kethineni, Sesha, & Cao, Ying. (2020). The rise in popularity of cryptocurrency and associated criminal activity. *International Criminal Justice Review 30* (3), 325–344, https://doi.org/10.1177/1057567719827051.
57. Atzori, Marcella. (2017). Blockchain technology and decentralized governance: Is the state still necessary? *Journal of Governance and Regulation 6* (1), 45–62, https://doi.org/10.22495/jgr_v6_i1_p5.
58. Danaher, John. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology 29* (3), 245–268, https://doi.org/10.1007/s13347-015-0211-1.
59. Bovens, "Public accountability."
60. Schumpeter, Joseph A. (2008). *Capitalism, socialism, and democracy*. 1st ed. Harper Perennial Modern Thought.
61. Christiano, Thomas. (1996). *The rule of the many: Fundamental issues in democratic theory*. Westview Press.
62. Kroll, "Accountability in computer systems."

References

Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). Algorithmic accountability for the public sector. August. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

Atzori, Marcella. (2017). Blockchain technology and decentralized governance: Is the state still necessary? *Journal of Governance and Regulation 6* (1), 45–62. https://doi.org/10.22495/jgr_v6_i1_p5.

Binns, Reuben. (2018). Algorithmic accountability and public reason. *Philosophy & Technology 31* (4), 543–556. https://doi.org/10.1007/s13347-017-0263-5.

Bovens, Mark. (2007). Public accountability. In Ewan Ferlie, Laurence E. Lynn, Jr., and Christopher Pollitt (Eds.), *The Oxford handbook of public administration* (182–208). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199226443.003.0009.

Bovens, Mark. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism." *West European Politics 33* (5), 946–967. https://doi.org/10.1080/01402382.2010.486119.

Bovens, Mark, Goodin, Robert E., & Schillemans, Thomas. (2014). Public accountability. In Mark Bovens, Robert E. Goodin, and Thomas Schillemans (Eds.), *The Oxford handbook of public accountability* (pp. 1–20). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0012.

Busuioc, Madalina. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review 81* (5), 825–836. https://doi.org/10.1111/puar.13293.

Christiano, Thomas. (1996). *The rule of the many: Fundamental issues in democratic theory*. Westview Press.

Danaher, John. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology 29* (3), 245–268. https://doi.org/10.1007/s13347-015-0211-1.

Dignum, Virginia. (2020). Responsibility and artificial intelligence. In Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 213–231). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.12.

Doshi-Velez, Finale, Kortz, Mason, Budish, Ryan, Bavitz, Chris, Gershman, Sam, O'Brien, David, Scott, Kate, Schieber, Stuart, Waldo, James, Weinberger, David, Weller, Adrian, & Wood, Alexandra. (2019). Accountability of AI under the law: The role of explanation. Preprint. *ArXiv:1711.01134 [Cs.AI]*, December 20. http://arxiv.org/abs/1711.01134.

Fernandez, J.D., & Vico, F.. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research 48* , 513–582. https://doi.org/10.1613/jair.3908.

Floridi, Luciano, & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines 14* (3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Foley, Sean, Karlsen, Jonathan R., & Putniņš, Tālis J. (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies 32* (5), 1798–1853. https://doi.org/10.1093/rfs/hhz015.

Goodin, Robert E. (2003). Democratic accountability: The distinctiveness of the third sector. *European Journal of Sociology 44* (3), 359–396. https://doi.org/10.1017/S0003975603001322.

Grant, Ruth W., & Keohane, Robert O. (2005). Accountability and abuses of power in world politics. *American Political Science Review 99* (1), 29–43. https://doi.org/10.1017/S0003055405051476.

Guynn, Jessica. (2015). Google photos labeled Black people "gorillas". *USA Today*, July 1.

Hutchinson, Ben, Smart, Andrew, Hanna, Alex, Denton, Emily, Greer, Christina, Kjartansson, Oddur, Barnes, Parker, & Mitchell, Margaret. (2020). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *ArXiv:2010.13561 [Cs]*, October 22. http://arxiv.org/abs/2010.13561.

Institute for the Future of Work. (2020, October). Mind the gap: How to fill the equality and AI accountability gap in an automated world. London.

Kethineni, Sesha, & Cao, Ying. (2020). The rise in popularity of cryptocurrency and associated criminal activity. *International Criminal Justice Review 30* (3), 325–344. https://doi.org/10.1177/1057567719827051.

Kohli, Nitin, Barreto, Renata, & Kroll, Joshua A. (2018). Translation tutorial: A shared lexicon for research and practice in human-centered software systems. *1st Conference on Fairness, Accountability, and Transparency*. February. New York.

Koppell, Jonathan G.S. (2005). Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public Administration Review 65* (1), 94–108. https://doi.org/10.1111/j.1540-6210.2005.00434.x.

Kroll, Joshua A. (2020). Accountability in computer systems. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 179–196). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.10.

Kroll, Joshua A., Barocas, Solon, Felten, Edward W., Reidenberg, Joel R., Robinson, David G., & Yu, Harlan. (2016). Accountable algorithms. *University of Pennsylvania Law Review 165* , 633.

Landemore, Hélène. (2020). *Open democracy: Reinventing popular rule for the twenty-first century*. Princeton University Press.

McGinnis, John O. (2010). Accelerating AI. *Northwestern University Law Review 104* (3), 1253–1269.

Mittelstadt, Brent. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1* (11), 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Morozov, Evgeny. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.

Mulgan, Richard. (2000). "Accountability": An ever-expanding concept? *Public Administration 78* (3), 555–573. https://doi.org/10.1111/1467-9299.00218.

New, Joshua, & Castro, Daniel. (2018). How policymakers can foster algorithmic accountability. Center for Data Innovation.

Nissenbaum, Helen. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2* (1), 25–42. https://doi.org/10.1007/BF02639315.

Nissenbaum, Helen. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Noorman, Merel. (2020). Computing and moral responsibility. In Edward N. Zalta (Ed.), *Stanford encyclopedia of philosophy*, Spring Edition. https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

Rahwan, Iyad. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology 20* (1), 5–14. https://doi.org/10.1007/s10676-017-9430-8.

Raji, Inioluwa Deborah, Smart, Andrew, White, Rebecca N., Mitchell, Margaret, Gebru, Timnit, Hutchinson, Ben, Smith-Loud, Jamila, Theron, Daniel, & Barnes, Parker. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. ACM, 2020. https://doi.org/10.1145/3351095.3372873.

Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1* (5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.

Schumpeter, Joseph A. (2008). *Capitalism, socialism, and democracy*. 1st ed. Harper Perennial Modern Thought.

Selbst, Andrew D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology 35* (1), 117–191.

Shoemaker, David. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics 121* (3), 602–632. https://doi.org/10.1086/659003.

Skelton, Sebastian Klovig. (2019). Accountability is the key to ethical artificial intelligence, experts say. *ComputerWeekly.Com*, December 16. https://www.computerweekly.com/feature/Accountability-is-the-key-to-ethical-artificial-intelligence-experts-say.

Skitka, Linda J., Mosier, Kathleen, & Burdick, Mark D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies 52* (4), 701–717. https://doi.org/10.1006/ijhc.1999.0349.

Sunstein, Cass R. (2003). Beyond the precautionary principle. *University of Pennsylvania Law Review 151* (3), 1003–1058. https://doi.org/10.2307/3312884.

Tutt, Andrew. (2017). An FDA for algorithms. *Administrative Law Review 69* (1), 83–123.

Wachter, Sandra, Mittelstadt, Brent, & Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics 2* (6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080.

Wagner, Ben. (2020). Algorithmic accountability: Towards accountable systems. In Giancarlo Frosio (Ed.), *The Oxford handbook of online intermediary liability* (pp. 678–688). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198837138.013.35.

Waldron, Jeremy. (2016). Accountability and insolence. In *Political political theory: Essays on institutions* (pp. 167–194). Harvard University Press.

Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics 24* (2): 227–248. https://doi.org/10.5840/philtopics199624222.

CHAPTER 9

GOVERNANCE VIA EXPLAINABILITY

DAVID DANKS

Introduction

Successful governance of AI systems requires some knowledge of how, and more importantly *why*, the system functions as it does. If I do not understand why a loan approval algorithm denied me a loan, then I cannot change things in hopes of a better outcome next time. If regulators do not understand why a self-driving car made a particular left turn, then they cannot safely determine when the car should or should not be used. If a doctor does not understand why a medical AI made a mistaken diagnosis, then she may not be able to provide feedback to improve its future performance. And similar observations could be made across a range of different domains and AI systems. In general, if relevant stakeholders—developers, users, citizens, and so forth—fail to understand why the system works as it does, then many aspects of governance either will be infeasible or will require costly trial-and-error.

This deeper understanding is particularly important in the context of AI because prior testing might not be feasible. One of the key motivations for deploying an AI system is that it is hopefully intelligent enough to adapt to novel, unexpected, or unforeseeable circumstances. That is, we want to use AI systems *precisely* when we cannot measure performance in all relevant contexts prior to deployment. Standard approaches to governance using mere reliability will almost always be insufficient for AI systems because they will inevitably be used in unexpected situations. Instead, we have even more reason to require an understanding of the functioning of the AI system—said differently, an understanding of why it works as it does.

This kind of understanding is, in everyday life, usually provided by explanations. I could ask a loan officer to explain my denial, regulators could ask human drivers why they made that left turn, or a doctor could explain her reasoning to a review board. So if we are interested in AI governance, then we might naturally be interested in AI systems that are, in some sense, suitable for explanations. Moreover, there has been significant research on these explainable AI (XAI) systems over several decades; the underlying technologies are

starting to mature. But while this connection between XAI and AI governance is intuitively appealing, matters are not so simple, precisely because explanations, XAI, and governance are all more complex than these initial observations suggest.

This chapter aims to show how different kinds of explainability can be used to support different functions of AI governance. While there is heterogeneity in explanations and XAI, there is a universal feature of all explanations that can be used to provide concrete guidance about how XAI can, and sometimes cannot, support AI governance. In particular, the quality of explanations depends on whether they support the goals of the explanation recipients. The result of this analysis is a concrete framework for "AI governance via explainability" or "explainability for AI governance," rather than specific policy or process recommendations. This framework is complex; there is no simple way to govern AI via explainability given its many different uses, contexts, and stakeholders. However, this complexity need not lead to paralysis; the motivating intuition is correct that explanations can sometimes improve AI governance.

This chapter has much less discussion on the topic of "explainability *through* governance" or "explainability as a goal of governance." There are many different methods and frameworks that have been proposed to help guide the development of XAI systems, some of which presumably count as forms of governance to reach the goal of XAI. At the same time, though, explainability is very rarely an end in itself; rather, explainability is a goal because it could help to increase use, performance, trust, control, or some other more fundamental goals that are central to successful (from a societal perspective) governance. Explainability through governance is important primarily because it can enable governance via explainability, and so the focus of this chapter will be on the latter possibility.

Section 2 explores the intuitive connection in more detail to show why explanations seem well-suited to support or enable governance, particularly for AI systems, but also why the details matter. Sections 3 and 4 then respectively consider XAI and explanations in more detail. Section 5 returns to the intuitive connection to develop the framework for AI governance via explainability. While AI governance cannot be guaranteed simply by using explainable AI systems, an appropriate use of explanations can significantly advance the governability of AI. Section 5 focuses on high-level considerations about governance without being tied to any particular proposed or current law, regulation, or policy. This agnosticism ensures that the framework can be used both to evaluate current ideas, and also to guide future ones. Section 6 provides some concluding thoughts. We begin, though, with a more careful consideration of the intuitive connection.

A *Prima Facie* Connection

As noted above, a salient feature of AI systems (in contrast with non-autonomous systems) is the difficulty of knowing or predicting how they will behave in novel situations or contexts. If the desired system performance could be fully specified ahead of time, then there would be no reason to use anything that might be called "AI" (rather than a much simpler algorithm or computational system), as one could simply "hard code" the desired behavior. Part of the usefulness of AI systems is precisely that they can be flexible and surprising, hopefully in positive ways. For example, successful self-driving cars must do more

than follow simple patterns, but rather adapt to the constantly changing roadways. And because AI systems will be used in novel or surprising situations, "mere" reliability information is insufficient for appropriate use and governance. The insufficiency of classical reliability information is exacerbated when, as frequently occurs, people have a very different understanding from the AI of which situations are novel.[1]

Instead, we need to know *why* the system behaves in particular ways. Explanations are often proposed to be answers to so-called why-questions, such as "why did she choose that career?" or "why did the bridge support fail?"[2] Hence, we might naturally look towards explainable AI (XAI) as more easily governable, or perhaps even the only kind of governable AI.

The field of explainable AI dates back several decades, and has experienced a renaissance in recent years. There are multiple kinds of AI that have been described as "explainable"; XAI is not one single technology. For example, a loan approval algorithm could be XAI if (a) it self-generates explanations of approve/reject decisions; (b) data scientists can analyze it to understand why it made particular approve/reject decisions; or (c) loan applicants can interpret it in ways that (seemingly) yield why-information. Of course, the same loan approval algorithm could fit multiple of these characterizations. XAI will be the focus of the next section; for now, we only need the observation that there are many kinds of XAI.

Given this diversity, AI governance via explainability will depend on a better understanding of the nature of explanation. Explanations do not merely *describe* some event or phenomenon, but rather provide (in some sense) an account of *why* it occurred. A theory of explanation must explicate what more is required for something to be an explanation, rather than a mere description. This explication could be either philosophical or psychological, depending on whether the focus is, respectively, what an explanation ought to be (normatively, rationally), or what an explanation actually is for humans (descriptively, cognitively).

Section 4 will consider theories of explanation in detail, but a high-level overview will be useful in the meantime. One classic philosophical theory of explanation is that explanations are simply predictions where the outcome is known;[3] that is, an explanation of some event is a description of the conditions from which one can predict that event (and we know that it actually occurred). In the context of AI, this idea would suggest that an XAI system would only have to provide the (relevant) input data that produced the output. This natural idea will not work without significant amendment, however, because description of the inputs provides only the initial conditions for the event, not the explanation of it. For example, a list of input symptoms will not provide an explanation for why a medical AI system diagnosed someone with a disease. More generally, a theory of explanations must account for the fact that some information conveys *why* something happened (i.e., is a genuine explanation) while other information only conveys *that* something happened or the conditions for it to happen.

Matters become even more complex with psychological theories of how people actually use explanations, or decide whether a description counts as an explanation. In particular, explanations might appear to be backward-directed, in the sense that they give a retrospective account of why something happened. Explanations seem, on the surface, to be about the past. However, it turns out that they have a significant forward-directed psychological function: explanations can help the recipient better predict and respond to similar situations in the future.[4] Psychologically, explanations are not just about what did happen in the past, but also about what might happen in the future. Whether considered from a philosophical or psychological perspective, explanations are complex objects.

This section has outlined a *prima facie* plausible argument: AI governance requires understanding "why?"; explanations answer why-questions; therefore, AI governance requires explainable AI. At the same time, even a high-level exploration of the steps in this argument has revealed complexities and nuances that are obscured by the *prima facie* formulation, and many more issues lurk just beneath the surface. We must dig deeper than this overly-quick two-premise argument. In particular, each different type of XAI requires a substantive theory of explanation in order to be usable. The features that are shared by different theories of explanation thereby determine what features must hold of any XAI system. That connection prompts the exploration of both philosophical and psychological theories of explanations to find those common elements.

A Taxonomy of Explainable AI (XAI)

The idea that an AI system might, or should, be explainable (in some sense) has a long history, dating back at least to the 1970s.[5] Many of those early examples were expert systems that were supposed to assist human decision-making, or perhaps replace human decision-making only after validation (partly) on the basis of expert knowledge. The desire for explanations was thus largely driven by skeptical humans who questioned the possibility that an AI system could help or replace them. That is, XAI was needed to convince humans that the AI system "knew what it was doing." XAI faded as a central topic with the rise of "big data" machine learning systems, including (but not limited to) advances such as deep neural networks. This rise changed the justification for AI systems to focus on their ability to identify patterns in data that were unnoticed or unlearnable by humans. And because part of the appeal of AI systems was exactly their ability to understand, predict, or control the world in ways that (seemingly) exceeded human powers, explainability no longer seemed so important. However, many recent events have highlighted the costs of this shift. For example, when AI systems understand the world in different ways than people, then it can be quite difficult to accurately predict or determine the contexts in which the AIs will fail. Non-explainable systems also typically do not provide useful insights that can be applied elsewhere; people instead must simply accept (or not) the AI system output. For these reasons (and many others), XAI has reemerged as a major topic of research in recent years.

The arguments in this chapter largely do not depend on technical details about XAI systems, so I will not provide a technical survey of the many different XAI methods. Interested readers should consult one of the many such surveys that are now available.[6] At the same time, I will use particular XAI approaches or techniques as examples (without much detail) in order to show how the topics in this chapter connect with that technical literature.

Any overview of XAI is complicated by the relative lack of agreement about terminology. There is a set of interconnected concepts and terms—explainability, intelligibility, interpretability, transparency, understanding—that are not used consistently across the field. For example, one person's "explainable AI" could be another's "interpretable AI." This section will not attempt to adjudicate those terminological disputes, but will rather focus on the different functional types that fall under the broad label of "XAI." As a result, some instances of XAI (according to this chapter) might be called something different in other

contexts. This terminological agnosticism will enable progress on the relationship between AI explanations and AI governance, and might even provide a new way to draw principled distinctions between different types of XAI.

With these caveats in mind, one can distinguish at a high level between three different, not mutually exclusive, types of XAI. These three types can blur together in some cases, but they provide a useful taxonomy for thinking about AI governance. The first type of XAI—what can be called *explanation-generating AI*—is one that is itself capable of providing an explanation of its behavior when queried or probed. For example, a loan approval algorithm that recommends approval for an applicant might provide an accompanying explanation for its judgment, such as the key counterfactuals about what changes would have led to rejection of the application.[7] These AI-generated explanations could potentially be generated by a sub-system that analyzes the original AI; this kind of add-on module for explanation can enable one to convert many different AI systems into XAI ones, potentially even in a post hoc manner.[8] The key characteristic for explanation-generating AI is that the system itself produces the explanation. The humans who develop, use, or otherwise interact with the AI need not do any particular cognitive work.

One challenge in assessing explanation-generating AI is that they are often expected to *justify* judgments, not merely explain them. For example, a loan approval algorithm of this type is sometimes expected to explain not only why it provided a particular judgment, but also why that judgment is legally, morally, or socially acceptable. There are three main reasons why this chapter will largely set aside the potential justification-generating capabilities of some XAIs, and instead focus on their explanation-generating capabilities and resulting implications for AI governance. First, explanation and justification are simply two different goals for an XAI system. In general, explanations purport to tell us why something happened (or did not happen) while justifications purport to say why something was right (or permissible or acceptable). That is, explanations are largely descriptive, while justifications involve a significant normative component that defends the action as appropriate. Second, there is often significant disagreement about which normative standard should be used, and so disagreement about how to evaluate an AI-generated justification. Third, the other two types of XAI systems are used much less frequently to try to produce justifications for system outputs, and so a focus on justifications would omit a large number of XAI systems.

Returning to the three types of XAI, the second type—*human-explainable AI*—arises when an appropriately knowledgeable or trained human is able to generate explanations, usually for themselves, of the AI behavior. Many canonical examples of XAI fall into this type. For example, shallow (i.e., few-layer) decision trees are widely thought to be explainable systems, but explanations of their output or behavior (e.g., "this person was approved for a loan because their credit score was high and their debt-to-income ratio was low") are actually generated by people reflecting on the model, not the AI systems themselves. A decision tree does not itself provide an explanation; the human plays an integral role in the production of an explanation for a decision. Other popular XAI techniques of this type aim to extract low-dimensional approximations of the actual high-dimensional model.[9] Again, the AI system does not itself generate any explanations, but rather provides information that is useful for a knowledgeable human who is trying to make sense of the system performance. Human-explainable AI is clearly dependent on the knowledge and skills of the relevant human. Most work has assumed that the relevant human is generally knowledgeable about

algorithms and computational models, and so can understand things like low-dimensional approximations without further training. At the same time, research on human-explainable AI usually requires that the required training is not *too* specialized, so some AI systems (e.g., deep neural networks) are consistently classified as not human-explainable even if a few people actually could generate an explanation from them.

Finally, a third type of XAI—*human-interpretable AI*—is one that exhibits patterns of behavior for which untrained people can generate satisfactory stories. These stories can "explain" the AI in the sense of capturing the patterns of AI behavior, while not necessarily counting as "explanations" on standard philosophical or psychological theories of explanation. Research on human-interpretable AI focuses on shaping or constraining the AI's behavior so that humans without specialized knowledge can produce "as if" stories, perhaps in the same ways that people generate stories about one another to explain behavior. For example, a robotic system is sometimes described as XAI if humans can understand the robot's behavior "as if" it had beliefs, desires, and other mental states.[10] Of course, those "explanations" might be entirely wrong about the actual inner workings of the robot; there might be no representations or content in the robot that fit our naïve understandings of beliefs, desires, and so forth. Nonetheless, if (untrained) people are able to generate stories that "make sense" of the AI behavior, then there is a sense in which the AI is explainable. At the least, these stories might enable people to achieve many of the goals that explanations normally support.

Each of these types of XAI could be useful in a particular context, and could improve or increase AI governance. One significant challenge, however, is that *each of these types of XAI requires a substantive theory of explanation*; the discussion in this section has taken for granted that we know what should or does count as an explanation. For example, consider explanation-generating AI: without a substantive theory of explanation, the developer would not know what kinds of explanations should be generated by the AI system, nor how to evaluate whether the AI system actually succeeds in achieving (this type of) explainability. Similar observations arise for the other two types of XAI, and so there is seemingly an explosion of types of XAI: the three here, multiplied by all of the different substantive theories of explanation. Such a proliferation of types poses a significant barrier to AI governance through AI explanation, as there may be simply too many different permutations. Successful governance requires a response to this challenge, but that response will require closer examination of the nature of explanations.

What Is an Explanation?

Both philosophical and psychological theories of explanation are relevant for potential AI governance. The former type of theory aims to articulate what an explanation ought to be, in some normative sense; what ought to be the features or characteristics of something that truly does answer a why-question? The latter type of theory characterizes how (purported) explanations actually function in human cognition; how do people use things that seem to be explanations in understanding, reasoning about, and acting in the world? And how do people determine that something might be an explanation? The philosophical and psychological theories can clearly diverge from one another: people might not correctly identify

and use "real" (according to some philosophical theory) explanations, and a philosophical theory might not make any reference to an explanation's cognitive impacts. Nonetheless, we should plausibly expect *some* connections between philosophical and psychological theories of explanation, much as we expect connections between such theories of causation, action, agency, and so forth. More importantly, both types of theories are relevant for governance via explainability, as both the information in an explanation (normative aspects) and people's responses to explanations (descriptive aspects) will impact AI governance.

Philosophical theories of explanation

At a high level, philosophical theories of explanation divide into two types—realist and pragmatic—depending on whether the quality of the proposed explanation is based on its accuracy or truthfulness, or instead based on its pragmatic value to the recipient of the proposed explanation. That is, these two kinds of theories differ based on whether explanations should mirror (in some sense) reality in the right ways, or whether they should support people's cognitive needs in the right ways. Many explanations will satisfy both requirements (truthfulness *and* helpfulness), but some explanations, including some kinds of XAI, might satisfy only one.

Realist philosophical theories of explanation hold (roughly) that an ideal explanation will articulate all-and-only the actual, true reasons why the explanandum—that is, the thing to be explained—occurred, perhaps in the particular way that it did. For these kinds of philosophical theories, a proposed explanation that gets the facts wrong is simply not an actual explanation, regardless of whatever other benefits might result from someone receiving it. For example, an explanation of why a tree's leaves are green should make reference to chlorophyll absorbing red and blue parts of the visible spectrum. In contrast, the proposal that magical fairies paint the leaves green when no one is watching would equally well enable correct prediction, generalization, and so forth, but would not be an explanation.

Of course, this simplistic characterization in terms of true facts cannot be the full story. In particular, a realist philosophical theory must provide the restrictions on a set of facts (about events, laws, causal relations, and so forth) that must hold for it to actually answer a why-question. For example, some accounts might require that an explanation include (necessary) laws of nature[11] or causal relations and structures,[12] or that it provide a unification of multiple events or phenomena,[13] or some other additional criteria beyond simply providing relevant facts about the events leading up to the explanandum. Moreover, explanations can sometimes seemingly include false-but-approximately-correct claims, as when one explains the changing tides by appeal to Newtonian gravitational forces (plus the changing location of the moon and other facts), rather than the laws of general relativity. An explanation thus must be the right (in a sense to be considered shortly) set of true facts, not just any arbitrary collection of true facts.

In contrast with realist theories of explanation, pragmatic accounts focus on the functional role that explanations ought to play for the recipient.[14] In general, explanations enable people to better understand what occurred, and pragmatic theories hold that this impact is the core characteristic of an explanation. That is, explanations are whatever increases understanding, even if it fails to mirrors reality. As a result, pragmatic theories allow for the possibility that false statements can nonetheless explain[15] (e.g., if they provide a useful analogy,

or a useful "as if" story as for human-interpretable AI). Moreover, understanding critically depends on the goals and/or context of the recipient, and so pragmatic theories of explanation argue that there is a component to every explanation that necessarily depends on features of the recipient. On a pragmatic account, the question "is this series of statements an explanation?" simply cannot be answered without knowing about the goals and/or context in which those statements are provided.

One obvious implication of a pragmatic theory of explanation is that the exact same statements could be an explanation for one individual but not for another.[16] For example, an explanation in terms of quantum mechanics might be useful for a physicist but not a young child; more relevantly here, an explanation in terms of a complex machine learning algorithm might be useful for an AI researcher but not a member of the general public. This audience-dependence is also endorsed by proponents of realist accounts of explanation, though primarily because of the pragmatics of conversation, not any necessary aspect of explanations themselves. That is, the proponent of a realist theory can acknowledge that we give different explanations to a child and a quantum mechanic, but reject the idea that this explanatory practice thereby tells us anything interesting about the nature of explanations.

The clearest point of departure between the types of theories is whether radically false (i.e., not even approximately true) statements can be part of an explanation: realists say "no" while pragmatists say "yes, if it increases understanding." Of course, radically false statements will often not contribute to understanding, so we should probably expect that most explanations will involve (approximately) true statements. Nonetheless, the question of whether radically false statements can *ever* be part of an explanation highlights the different grounds for explanations—accuracy vs. understanding. This question is particularly salient for human-interpretable AI systems because the stories that people generate might involve exactly these kinds of radically false statements (e.g., "the robotic car *believes* that there are people in the road"). This question is also salient when people look to explanation-generating AI systems for justifications because justifications are rarely evaluated based on their helpfulness.

As noted above, the diversity of normative theories of explanation potentially poses a challenge for AI governance because there could be a problematic proliferation of XAIs. However, if there are features or properties that are shared by (almost) all substantive theories of explanation, then those can be used for AI governance via explainability, regardless of the particular type of XAI. One can remain agnostic about which normative account is right, and instead simply use the shared features and properties. The resulting methods and practices would have force and legitimacy across a wide range of settings, commitments, and approaches. Agnosticism about the "true" nature of explanation can thus be seen as analogous to agnosticism about "the good life" that underlies many governance systems for political life (in value-pluralistic societies).

One might reasonably wonder whether there are *any* features that are shared by all substantive normative theories of explanation. I propose that the (broadly understood) goals of the explanation recipients are necessarily relevant for each of these types of normative theories. Obviously, the recipients' goals are critically important for pragmatic theories; one cannot know whether something contributes to understanding without knowing why the recipient wants to understand. In contrast, realist accounts seemingly make no explicit reference to goals, but I contend that they involve an implicit dependence on goals.

In particular, recall that realist accounts must be supplemented in some way to indicate which sets of (approximately) true statements constitute an explanation. This addition could be provision of a measure for "approximate" truth, or restriction to certain sets of causal relations, or specification of neighboring theories that are unified via the proposed explanation, or many other supplements. In each case, though, the justification of a restriction will depend on the (broadly understood) goals of the recipient. For example, a restriction to specific causal relations might be appropriate only if the goal is control (which requires causal knowledge). Or what counts as an acceptable level of approximation will depend on goal-specific features. These implicit goals could be incredibly broad such as "know more about the world," but even that goal still contrasts with other possible goals (e.g., "better control the world"). Moreover, these goals are not tied to particular levels of description;[17] this goal-dependence is *not* an instance of the previous observation that conversational pragmatics can influence what explanations we happen to offer. Rather, full specification of a realist theory of explanation requires (implicit) specification of the recipients' goals in order to ground or justify the necessary, additional constraints on (approximately) true statements. Hence, we can see goal-dependence as a shared feature of substantive philosophical theories of explanation.

Descriptive theories of explanation

Now consider descriptive theories of explanation: what role do explanations play in human cognition? If explanations are to improve AI governance, then their use should depend on how the cognition of relevant stakeholders (e.g., developers and users) is influenced by those explanations. Of course, people's cognition will change after receiving *any* set of statements; for example, if I read some statements that I think are true, then I will have new beliefs, additional inferences from those new beliefs, and so on. The challenge for psychological theories of explanation is to determine what additional cognitive changes result when one receives an *explanation*, not just a set of statements.

One change that is not particularly relevant is people's subjective experience of liking (or not) a putative explanation. Cognitive changes and the phenomenology of explanations could presumably separate: something could provide cognitive benefit without people liking it, and vice versa (as is frequently found in pedagogical studies, or when people like something solely because it is familiar). For the purposes of AI governance, the cognitive impacts are the most relevant, rather than the experiential ones. Explanations can presumably provide a useful mechanism of AI governance only through changes in people's subsequent decisions and reasoning, rather than through a momentary good or bad experience (though with the caveat that a sufficiently bad subjective experience might lead someone to ignore the explanation). Whether someone "likes" an explanation—or even is willing to call something an "explanation"—is not the focus here; the question is how people think and decide differently as a result of the explanation.

One important feature of the cognitive impact of explanations is that they alter, hopefully for the better, people's *future* reasoning, prediction, and action, not only their knowledge of the past.[18] The statements in an explanation almost always refer to past features of the world, including both the past state of the world and the scientific laws and causal structures in place at the time. If one is provided this set of statements in a non-explanatory context (e.g.,

if the statements are a mere description), then one's cognition about the past will change as something is learned, but cognition about the future will not significantly shift. If these statements are instead presented as an *explanation*, then numerous studies have shown that one's cognition about the future will also change.[19] For example, suppose I see a fallen tree and am told "there was a beetle infestation last year." If this claim is presented as merely a description of the forest, then I simply learn about some events from last year. If that statement is instead presented as an explanation, then I infer more, such as that beetles are the kinds of things that can lead to fallen trees. Future predictions will change in light of this new knowledge in ways that go beyond the impact of the facts about last year.

The future-directed impacts of explanations can be understood in terms of generalization. Explanations indicate the features of the world that are relevant to understanding why something occurred, and so convey information about which features are likely to be relevant in future contexts. When the fallen tree is explained in terms of a beetle infestation, then if I care in the future about predicting or preventing dead trees, then I should seek information about beetle infestations. While various descriptive theories might differ about the exact impacts on future cognition, they share the conclusion that explanations are not purely backward-directed but have significant forward-looking impacts.

As this example shows, the descriptive quality of an explanation will depend on whether it enables the right kinds of future cognition. That is, whether something is a good explanation (in descriptive terms) will depend on whether it provides the information for the recipient to succeed at relevant future cognitive tasks. But the relevant future cognition will depend on the goals and needs of the recipient: something could be a good explanation for certain goals, but if I never actually encounter those corresponding cognitive tasks in the future, then it is not helpful for my particular cognition (and so not actually a good-for-me explanation). These goals could be quite broad and vague (e.g., "be prepared for surprises in the future"), but the psychological quality of an explanation nonetheless depends on them.

Goal-dependence or -sensitivity thus emerges as one universal feature, perhaps one of many, across essentially all substantive theories of explanation, whether philosophical or psychological, though the details of that dependence can vary. The next section shows how to use this universal feature to better understand how XAI might, and might not, be used to improve AI governance. If there are other universal features of substantive theories of explanation, then those could also be incorporated in similar ways.

Governable AI via Explainability

We start by considering some of the goals of AI governance, as those will constrain the type of explainability that might be useful for governance. In particular, the goals of AI governance might require certain kinds of explanations, and thus certain kinds of XAI, at least to the extent that we care about governance via explainability.[20] I adopt a relatively general notion of governance as the mechanisms that steer the governed towards desired outcomes and targets, similar to a forward-looking version of notions such as "accountability as a practice."[21] Governance on this broad conception has the overall function of providing some level of assurance that our AI systems will bring about the outcomes we want, and also that appropriate responses will be taken when they fail to do so. In this

section, I consider the implications for XAI of four different potential requirements for this type of broad AI governance: system prediction, system control, failure signals, and proper incentives. Of course, these four form only a partial list; no claims of completeness are made or intended, though these four features will arguably be relevant for any governance process. How should these requirements, and the corresponding goals to achieve each, constrain the types of XAI that might be developed or deployed?[22]

The first requirement was mentioned at the start of this chapter: making predictions about system performance in novel circumstances. Appropriate governance mechanisms require the capability to make (noisy, defeasible) inferences about the likely AI performance in new situations so that the appropriate contexts or scopes for its use can be determined. Prediction for novel circumstances is critical to address this governance challenge. Explanations can clearly support predictions in novel circumstances, but they need to be either realist explanations or pragmatic ones with this goal. Explanation-generating and human-explainable AI systems are thus likely to be helpful. In contrast, human-interpretable XAI systems are typically built so that people can construct stories for normal operation, and those stories will not necessarily provide accurate predictions in novel contexts (whether because the stories are not accurate, or they have the wrong pragmatic goal). For example, I might interpret a robot as if it has human-like beliefs and desires, only to be quite surprised at its behavior in new situations if it does not *actually* have beliefs and desires. Regardless of the type of XAI, it should lead to explanations (or stories) that prioritize the goals of system deployers. This requirement is needed for governance over contexts of use, and deployers are the individuals who have the largest impact on that aspect of AI systems. Explanations that instead help users make predictions, for example, would not necessarily support this governance function because users have relatively little control over deployment contexts.

A second goal is making predictions given interventions or changes to the AI system, relevant contexts, or human users. Governance requires mechanisms that can shift the behavior of the governed system in appropriate ways, which presupposes some ability to estimate how the system might respond to such changes. Governance mechanisms should only prescribe various adjustments to an AI system given reasonable inferences about the results of such changes. Prediction given interventions is importantly different from prediction given observations. One can predict that the current temperature outside is cold by observing people wearing heavy jackets, but intervening to force people to wear heavy jackets in summer will not lower the temperature. Both kinds of prediction—from observations and from interventions—are important for the design and use of successful AI governance, but they must be separately supported. Similarly to the first goal, explanation-generating and human-explainable AI systems are likely to be helpful,[23] but human-interpretable AI systems will not necessarily provide appropriate explanations for this requirement, unless those stories happen to correspond to the actual causal structure of the AI system. In contrast with the first requirement, these explanations (or stories) should be appropriate for both deployers and users, as both are likely to be in a position to change or impact the AI system.

The third governance requirement is knowledge or understanding of indicators of failure or problems, as this awareness is a prerequisite for appropriate monitoring and oversight. AI systems deployed in open contexts will inevitably surprise us, whether in good or bad ways. Their full performance profile will almost never be known in advance of their use, and so governance requires mechanisms to detect problematic AI behavior. Hence, governance

via explainability should support the goal of appropriate detection capabilities, where this goal is shared by both regulators and users. One key criterion for fault detection is to distinguish between errors that should be corrected and the inevitable failures that are simply part of normal operation in a noisy world. For example, a loan approval algorithm will surely not be perfect, but its failure can have different sources. Some of its judgments will be wrong simply because they are based on imperfect, partial data, while others might be wrong because of systematic (and legally problematic) biases in the algorithm. A good governance mechanism should minimize or mitigate the latter kind of errors, but that requires the ability to distinguish between these kinds of errors. For this requirement, explanation-generating and human-explainable AI systems can provide the required information for regulators and users. More interestingly, human-interpretable AI systems can also provide useful stories, though only if those stories are tied to the identification of appropriate behavior. Human-interpretable AI systems will not necessarily enable one to know how to respond to failures, but they can help to identify those failures.[24]

The fourth requirement for AI governance extends out from the technology to include the humans involved in its design, development, deployment, and use. In particular, people will frequently be "in the loop" with AI systems, and so those people's actions must be taken into consideration when aiming for governance. Even a well-designed AI system could lead to problematic outcomes if people deliberately misuse it. For example, racist people using an unbiased loan approval algorithm could do a great deal of harm that proper governance should minimize or mitigate, but the focus should be the people not the AI. However, governance mechanisms will typically not be able to constantly monitor the people, and so successful governance requires the creation, implementation, and maintenance of proper incentives to ensure appropriate behavior. One might note here that XAI does not seem particularly relevant to proper incentives, and that observation is exactly the point. This particular requirement for AI governance is included precisely because the move from AI to XAI does not advance it in any substantive way. Increased understanding of the AI system will probably not help to understand or create proper incentives. Explanations of the AI system, regardless of type or source, will simply not help one to understand how the broader social system could (or should) be changed, particularly when there are significant systemic biases in our data or society. Although governability can be improved via explainability, XAI is not a panacea for all challenges of AI governance.

Conclusions

As AI systems proliferate in number, authority, and autonomy, there is increasing need for mechanisms to govern them in various ways. Explainable AI superficially holds the promise to enable the necessary governance; one might even be tempted to require all AI to be XAI in order to ensure that the systems are governable. This temptation is understandable, but also ultimately misguided. As demonstrated in this chapter, AI governance via explainability is a complex possibility that depends on the type of XAI, type of explanation, and relevant requirements or goals of the governance effort. At the same time, this complexity need not be overwhelming or paralyzing: there are commonalities within each of

these dimensions that can enable us to provide concrete guidance, primarily shaped by the goals and needs of the recipients of the explanations.

This chapter has provided a framework for pursuing AI governance via explainability, but it clearly has not been exhaustive. For example, one reason to have a governance system is to increase trust and utilization of a system: if one knows that there are mechanisms in place to nudge the AI towards better (in some sense) behaviors, then one is more likely to trust, and therefore use, that system. Explanations and XAI can potentially increase trust,[25] though it is an open question whether they do so in ways that support governance, or whether there are routes other than explainability to build the appropriate trust.[26] Nonetheless, the observations and arguments in this chapter can provide a schema for determining the ways in which explanations of various types do, or do not, support this potential goal of AI governance.

More generally, some high-level observations are in order. First, it is highly unlikely that *any* single AI system could exhibit explainability for all governance goals. Different aspects of AI governance connect with different roles, different goals, and different contexts. The exact same code or algorithm might be appropriate for one role, goal, and context, but not for another. That is, XAI should not really be understood as a type of AI, but rather as a type of AI-individual-society hybrid system, and so efforts at governance via explainability must also have this broader focus. Second, we need to think carefully about whose needs and interests are relevant for a particular governance function, as explanations must be tied to what people actually need and know, rather than an AI researcher's guesses or biases about those. Too frequently, XAI systems are built using the developer's beliefs about what will help deployers, users, or regulators, but without any serious effort to test or confirm those beliefs. One practical response would be to embrace the many calls for increased diversity and participation in the design, development, and deployment of AI systems, as those can help developers better understand the explanation needs of others.

Third, and perhaps most challenging, there is a deep tension between a system being widely interpretable, and it being appropriately governed. The human-interpretable type of XAI—observers can generate a story—is increasingly widespread, particularly through anthropomorphic representations of AI systems. For example, digital assistants (e.g., Siri, Alexa) are designed to help users generate stories about what those assistants "know" or "want," even though those stories are often incorrect. These systems can be interpreted, and "explanations" generated about them, even by people who have no technological training. This kind of XAI is thus particularly appealing for technologies that will be widely deployed, particularly because people arguably need understanding to freely consent to using such systems. However, the previous section showed the ways in which human-interpretable AI is less appropriate than the other two at supporting a wide range of governance functions. Because the human-generated stories need not be grounded in the underlying mechanisms or informational-causal structure of the AI system, they will inevitably fall short for those governance functions that depend on deeper understanding. There is thus an important, unresolved tension that will need to be resolved in coming years: widespread explainability (i.e., most people can generate a story) is insufficient for widespread governance (i.e., systems that are widely deployed and used), but we ultimately require both.

Acknowledgments

Thanks to Cameron Buckner, Jon Herington, Johannes Himmelreich, Ted Lechterman, Juri Viehoff, and Kate Vredenburgh for valuable feedback on earlier versions of this chapter. Significant portions of this chapter were written while the author was on the faculty at Carnegie Mellon University.

Notes

1. Though we should be thoughtful about whether we are, or should be, requiring more from our AI systems than we expect from human decision-makers; see Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology 32* (4), 661–683.
2. Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh Press; Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
3. Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science 15* , 135–175.
4. Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition 99* , 167–204.
5. Some notable early papers include: Clancey, W. J. (1983). The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence 20* , 215–251; Scott, A. C., Clancey, W. J., Davis, R., & Shortliffe, E. H. (1977). Explanation capabilities of production-based consultation systems. *American Journal of Computational Linguistics*, 1–50; Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence 21* (3), 285–325.
6. There are *many* different introductions and overview of XAI. Some high-level surveys include:

   Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access 6* , 52138–52160; Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics 4* (37).

7. Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI) 8* (1), 8–13); Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems 34* (6), 14–23.
8. Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems 33* (3), 87–95.
9. Examples include: Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777); Plumb, G., Molitor, D., & Talwalkar, A. (2018, December). Model agnostic supervised local explanations. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 2520–2529); Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

10. Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology 8* , 1962.
11. Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science 15* , 135–175.
12. Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press; Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
13. Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy 71* (1), 5–19; Kitcher, P. (1981). Explanatory unification. *Philosophy of science 48* (4), 507–531; Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. C. Salmon (Eds.), *Scientific explanation* (pp. 410–505). University of Minnesota Press.
14. Achinstein, P. (1983). *The nature of explanation*. Oxford University Press; Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
15. Bokulich, A. (2011). How scientific models can explain. *Synthese 180* (1), 33–45.
16. Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science 83* (5), 721–732.
17. Danks, D. (2015). Goal-dependence in (scientific) ontology. *Synthese 192* , 3601–3616.
18. Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences 10* (10), 464–470; Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition 99* (2), 167–204.
19. For an overview, see Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences 20* (10), 748–759.
20. See also Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence 296* , 103473.
21. Lechterman, T. M. (this volume). The concept of accountability in AI ethics and governance. *Oxford Handbook on AI Governance*. Oxford University Press.
22. A related question is asked by Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 1–24.
23. Though one must be careful to ensure that the proposed actions are actually feasible; see Barocas, S., Selbst, A. D., & Raghavan, M. (2020, January). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80–89).
24. This connection between explainability and ability to determine which failures are "reasonable" has also been discussed by: Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence 2* (12), 731–736; Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science 87* (4), 568–589.
25. Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems 20* (6), 542–556; Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305).
26. London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report 49* (1), 15–21.

CHAPTER 10

POWER AND AI
Nature and Justification

SETH LAZAR

Introduction

During the last decade, artificial intelligence (AI) has come to have ever more direct impacts on the lives of ordinary citizens and consumers. In particular, it is increasingly being used by some to exercise power over others. And it is enabling those with power to exercise much *more* power than they could without it. In response, scholars working on the societal impacts of AI and related technologies have advocated shifting attention from the question of how to make AI systems beneficial or fair towards a critical analysis of these new power relations.[1] But what normative lessons should we draw from these new analyses? Should the fact that AI systems are used to exercise power be cause for moral concern? In this chapter, I introduce the basic conceptual materials with which to formulate these questions and offer some preliminary answers.

I first introduce the concept of power, highlighting different species in the genus, and focusing on one in particular—the ability that some agents have to shape others' lives. I then motivate a theory of this species of power and answer some preliminary objections. Next, I explore how AI enables and intensifies the exercise of power so understood, followed by a brief sketch of three normative problems with power with suggestions of three complementary ways to solve those problems. I start, though, with one definition, and three caveats.

AI is a set of technologies enabling computer systems to perform functions like making inferences from data, optimizing for the satisfaction of goals within constraints, and learning from past behaviour.[2] Machine learning (ML) is a field of AI, using algorithms, statistical models, and recursive simulation to detect patterns in massive datasets, enabling computers to perform these functions without following explicit instructions. Advances in computational power and an explosion in the availability of training data have led to progress in ML that has made AI systems startlingly effective at these tasks, enabling them to play a pervasive role in our lives.

First caveat: definitions are needed so that we don't talk past each other, but very little turns on whether some particular computational system is properly described as "AI." If it involves computation and some measure of automation, then many of the same ethical and

political issues arise whether it's ultimately powered by deep neural nets, by simple logistical regression, or even by old-fashioned symbolic reasoning. As one way to see this, consider early papers by Roger Clarke, Helen Nissenbaum, and Danielle Citron, all of which antedate contemporary progress in ML, but which describe moral and political problems of computing systems—"dataveillance," diffuse accountability, power without due process—that remain entirely relevant today.[3]

Second caveat: in this chapter I explore how AI enables some people to exercise power over others, with a focus on facilitating the normative evaluation of computational systems. This is an introduction to power from the perspective of political philosophy, not an introduction to the methodology of the political economy of AI. The former task is undoubtedly relevant to the latter, but the political economy of AI draws on a much larger and more varied conceptual toolkit than I can feasibly introduce here.

Third, related caveat: within political philosophy, I advance an approach grounded in normative analytical political theory. Few concepts in the social sciences and humanities have been subjected to more dispute than this one; I cannot hope to either relitigate or faithfully represent those many controversies in one chapter, and many other resources are available for that end.[4] Instead, I take one approach, and apply it to the case of AI.

Power Over and Power To

The first task, then, is to fix what we mean by power for the purposes of this chapter. And the first key distinction is between "power to" and "power over."[5] On the first approach, power is a resource distributed around society—some people, or social groups, have more power than others, and we can think of it in terms of what individuals or groups have power to do.[6] This can be understood very expansively—so that power to is really just a way of describing ability;[7] or more narrowly, for example by considering different social groups' ability to affect collective decision-making.[8] We might say, for example, that wealthier voters have disproportionate power to influence political outcomes compared to low-income voters.

On the second approach, power over describes a social relation whereby some entity has power over some other entity. For A to have power over B means, minimally, that A can get B to do things that B wouldn't otherwise do.[9]

Most scholars who emphasize the importance of thinking about AI and power are thinking about power over. But AI has significant implications for power to, as well. Think, for example, of the ways in which AI, robotics, and related technologies increase our capacities to get things done. And the power to that AI generates is unevenly distributed—it empowers some more than others. Most obviously, it empowers those who have access to it, which is in general those who are already advantaged. In other words, AI increases the power to bring about desired results of those who are already privileged. This is of obvious moral concern.

We ought not disregard power to. However, understood in this way, we *can* assimilate it to other concepts used in a theory of distributive justice. Power over, as we will see later, raises somewhat distinctive normative questions, with particular urgency in the age of AI.

Power over is a two-place relation: A has power over B. Our first task is to populate those two places. The simplest example involves one person having power over another person.

Power can also be exercised over other entities, such as animals, but I will focus on interpersonal power.

Individuals, groups, and aggregates can have power over individuals, groups, or aggregates. By "group" I mean a set of individuals with some internal structure—for example, a set of rules for membership or for decision-making. By "aggregate," I mean a set of individuals or groups without such an internal structure.

On a different understanding of power over, people can not only be subject to the power of agents, but also to that of social structures.[10] Social structures are (roughly) networks of roles, relationships, incentives, norms, and cultural schemas (widely shared sets of evaluative and doxastic attitudes), which can be populated or observed by different people at different times, which are generally the emergent result of human interaction over time, and which reliably pattern outcomes for people who are within or affected by them.[11] The presence of a particular social structure increases the probability that people who meet a particular description will experience a particular kind of outcome. Indeed, social structures can clearly have effects on people which, if they were traceable to some individual's decision, would lead us to say that that individual has power over those people. So, do social structures have power over us?

This question has exercised social theorists for decades, and we cannot hope to settle it here. The concept of power is, I think, appropriate for both purposes. However, in my view power is a social relation between people within society. Social structures often shape that social relation, by giving some people power over others, but all power is ultimately exercised by people. Moreover, the distributional effects of social structures on people's prospects and opportunities can be adequately described with other concepts, whereas the social and agential relation of power over cannot be described in any other way. I note, however, that a theory of agential power definitely demands a theory of the political affordances of social structures, which often clearly determine whether and to what extent one person has power over another.[12] Think, for example, of the way in which social structures of law, policing, and criminal justice in the United States give White Americans power over Black Americans.

Can we be subject to the power of only human agents? Could we be subject to the power of computing systems? If an automated system can effect a state change that would amount to an exercise of power if done by a person, that certainly *looks* like agential power. But if the system is simply deploying a set of pre-programmed rules, it might be only a tool for the exercise of power, rather than itself exercising power—we're *really* subject to the people who put the system in place over us. If the system's programming instead derives from its own learning, then perhaps the computing system itself exercises power.

The Content of a Theory of Agential Power

Very roughly speaking, power is about getting others to do things they wouldn't otherwise have done. Some think that power just consists in its actual exercise.[13] Others think power over is an ability (and so, on some views, a subset of power to). A has power over B just

in case A is able to get B to do something they would not otherwise do.[14] On my view, the latter approach is more fruitful, but some theorists might prefer to call the ability to exercise power "potential power."

More important, is that rough definition right? Is power just about getting people to do things they wouldn't otherwise do?[15] While this is clearly the paradigmatic case of power, it cannot ground a complete account, for at least three reasons.

First, it makes A's power over B a function of B's stoicism. Suppose A commands B to perform some act on pain of death, and B refuses. Then A has not exercised power over B even if A kills B for their disobedience. If C obeys the same command under the same circumstances, then A has power over C but not over B. Similarly, suppose that A can in fact do very little to affect D, but D is especially craven, and so is willing to do anything A says rather than suffer the merest expression of A's disfavor. Then A has even more power over D than over C and B. Although this is a controversial point, many will find this counterintuitive—A's power over B, C, and D should at least in part be a function of the objective degree to which A can affect their interests, not only of their purely subjective disposition to change their behaviour at A's command.

Second, and more generally, this interpretation of the content of power rests heavily on our being able to populate the counterfactual of what B would otherwise have done.[16] This might be relatively simple for discrete instances of the exercise of power. But consider the power that a state has over its citizens. How can we assess whether the state can make us do things we would not otherwise do? Must we imagine what we would have done if the state didn't exist, to establish the state's power over us? Or if the state were simply different in some way? Or if the state had issued some other command? To establish what B would have done absent A's intervention, we need some settled and neutral baseline to compare that intervention against. When A is sufficiently imbricated into the circumstances of B's life, it can be impossible to establish what that baseline would be.

Third, imagine that the gods on Mount Olympus have power of life and death over us, but have no way of communicating with us. In one sense, then, they lack power over us, because they cannot bring us to do anything that we would not otherwise have done. But they have power of life and death over us! That is real power, even if it is not used to shape people's choices.

A's getting B to do something they would not otherwise have done is paradigmatic but not definitional of power. A can also exercise power over B just by affecting B's interests—that is, by harming or benefiting B. The mute deity exercises power over us mere mortals. The stoical rebel is still under the power of the extortionist who can repay their stoicism with death.

A's agential power over B consists, then, in A's ability to make decisions that affect B's interests or choices. But is this definition overinclusive? Can't we all, all the time, make decisions that affect one another's interests or choices? Are we all therefore subject to one another's power? A further qualification is clearly necessary.

If A's power over B is matched by B's power over A, then neither has power over the other. Power over consists in an *asymmetry* between A and B—A can do something to B, and B cannot reciprocate in any comparable way. But it is not enough to simply describe power as being non-reciprocal. If A can significantly affect B's interests, but faces serious adverse consequences from C if he does so, then A does not have power over B. A has power over B, then, just in case A is able to make decisions that affect B's interests or choices without facing comparably adverse consequences.

Do all effects on B's interests and choices count? What if A can make B's life go much better by giving B $1,000,000? In my view, A's ability to do this gives them power over B. Does this mean, then, that billionaires have power over all of us whose lives would be substantially affected by a gift of such magnitude? No doubt the mere fact of extreme wealth *does* give some power over others. But there are compossibility constraints on power—Elon Musk and Jeff Bezos might be able to change any given individual's life at the stroke of a pen without adverse consequences resulting, but, as exorbitant as their fortunes are, they could only actually affect a relatively modest number of individuals to this degree. And of course, they have no idea who most of us are, so it is implausible to say that they are *able to* influence our choices or interests. But yes, billionaires do have too much power over others—that's one reason for wealth taxes that make such extreme wealth impossible.

How AI is Used to Exercise Power

With this working definition of power in hand, we can start to explore how we are subject to the power of AI and related computational systems in many different domains of our lives. I'll consider three main modalities for the exercise of power: intervening on people's interests, shaping their options, and shaping their beliefs and desires.

We can start with effects on the subject's interests.[17] AI systems are frequently used to support the allocation of resources within a population. This amounts to the exercise of power by the decision-maker (the individuals or organizations making use of the AI decision-support tool) over those who either benefit or don't from that decision. In the public sector, think of the allocation of healthcare, housing, social welfare resources, or algorithmic allocation of visas.[18] In the private sector, think of insurance pricing, decisions over whether a loan is granted, or automated systems that determine what products, services, and content you are exposed to online.[19]

What AI gives, AI can also take away: AI is used to allocate harms, as well as to directly harm people. The use of AI in the criminal justice system to inform decisions over pre-trial detention, sentencing, and parole is an obvious example where AI is used to exercise one of the most profound and serious powers that the state holds over its citizens.[20] The broader role of AI in policing—from facial recognition to predictive algorithms used to allocate police resources—typically meets this description as well.[21] But AI is also used to surveil populations—in workplaces by employers, in society at large by the state—which is a direct harm, and another way in which AI is used to exercise power.[22]

AI and related computational systems also shape people's choices by intervening on their options. For example, they can directly determine people's choice sets, making some options unavailable. This is particularly common in our digital lives, where computational systems dynamically adapt to make some things possible, others impossible. This is described by Roger Brownsword as "technological management"—the practice of shaping people's behaviour not by penalties, but by making undesired behaviour technologically impossible.[23] Of course, computational systems using AI can also add penalties to dispreferred options, as well as obstruct them or make them harder to perform, as with the "dark patterns" that make controlling one's privacy online so challenging.

As well as dynamically removing options, computational tools can dynamically *create* options. This too is a kind of power, roughly analogous to what Foucault called "governmentality."[24] For example, consider the simple practice of suggesting people to "friend" or "follow" on social media. This can create opportunities for new kinds of social interactions.[25] Or consider the personalized delivery of advertisements. Although sometimes a pure nuisance, these ads may involve presenting people with opportunities or options that are not otherwise easily accessible: for example, an investment opportunity, or a job posting, or even drawing the user's attention to a new area where they might consider purchasing a home.[26] At the intersection of both examples, consider the prospective role of AI in developing the "metaverse," at the intersection of online and offline lives. Companies like Meta (formerly Facebook) are now establishing the options that will be available to us in this new domain, and AI and related tools will shape what is possible for us within it—making some things impossible, others possible.

Besides actually creating new options, AI and related technologies have given "nudge" economics a profound shot in the arm, enabling (sometimes crude) ideas from behavioural economics to be operationalized at massive scale, with the ability to do massive social experiments enabling persistent fine-tuning.[27] AI is first used to profile individuals to tailor particular messages to them, then to deliver the right message for that person (as well as to manage the auction where that message competes with others for their attention), and then to learn from that trial in order to refine the message or select among alternatives.[28] Karen Yeung has described this as "hypernudging," but it is really more of a shove than a nudge and it has spawned an entire new field of "persuasive technology".[29]

Nudging (hyper or otherwise) is supposed to focus mostly on how people's options are presented to them. But hyperpersonalized computational systems are also well suited to more directly shaping people's beliefs and desires. Sometimes this is transparent, and well-intentioned. At other times it is both deceptive and extractive. In either case, it involves the exercise of power. We can distinguish roughly between three modalities: when AI is used to connect us to (hopefully) authoritative sources of information; when it mediates horizontal communication between users of the internet; and when it is used to identify and target us for specific persuasive messaging. In each case, AI systems shape our beliefs and desires, and the people who design these systems thereby exercise significant power over those affected by them. Indeed, search and recommender algorithms deploy, and have driven research in, some of the most advanced techniques in AI, using deep neural networks, large language models, and reinforcement learning among others.

Start with AI's role in shaping what we know by giving us access to (hopefully) authoritative sources of information—for example in helping us find relevant public health information during the COVID-19 pandemic. Many of us are practically dependent on digital platforms to guide us to this information. The platforms' intentions seem to be broadly good—they want to help us to find authoritative sources—but they still must make many controversial choices about what to show, what to exclude, and what to prioritize. There is no "neutral" path, especially given deep and persistent disagreement among people as to what information matters. They cannot avoid exercising power. And given the structure of our digital information environment, as well as the volume of information available, they must build these decisions into the search algorithms on which we rely to navigate the functionally infinite internet.

Recommender systems mediate horizontal communication among internet users—determining whose speech (and posts, engagements) is removed, muted, published, or amplified. Although human curation and content moderation plays a significant role, substantial automation is unavoidable given the sheer volume of information at stake.[30] These systems often rely on algorithms that can optimize for some measurable feature, which may only be an inadequate proxy for the properties that really matter. Engineers must choose not only what to aim at—what kind of information economy they want to achieve—but also how to operationalize that objective by choosing some measurable objective function. And the stakes are high—although researchers have long argued that optimizing for user engagement generates adverse social impacts, we now know that this has also long been verified by internal researchers at Meta, for example.[31] Prioritizing content that generates user engagement has led to the spread of radicalizing misinformation, shaping people's beliefs and desires in deeply harmful ways.

As well as connecting users to authoritative sources, and to each other, the same basic tools are used to connect businesses to users to extract value from the latter. The goal here is clearly on the borderline between persuasion and manipulation.[32] Online behavioural advertising is sometimes perfectly transparent and non-deceptive—it's simply about showing you a product that you might be interested in, given contextual cues from your entry in a search engine or the website you are visiting. But often it is not. You are being shown this advertisement because a profile of you has been built up from digital breadcrumbs left across multiple apps and websites, as well as edge computing devices.[33] You are receiving *this version* of the advertisement because automated testing over massive populations (none of whom knew they were test subjects) discovered that this wording worked best for people like you. In the extreme scenario, your personality type or your susceptibility to a particular persuasion technique is being automatically operationalized to increase the chance that you will be persuaded. And while all these different measures might make relatively little difference to the probability that *you* will buy the advertised item, over the whole population it does make a difference. Even if persuasive technologies are relatively unsuccessful at manipulating individuals, they enable an accelerated version of "stochastic manipulation" of populations at large.[34]

How we see the world is profoundly mediated by our digital infrastructure, and AI is integral to that infrastructure. Choices made by the designers, developers, and deployers of AI systems determine how the world is represented to us. Even were these decisions not made with any particular intention to shape people's behaviour one way or another, the ability to structure how billions of people perceive the world, through search algorithms and recommender systems, involves an extraordinary level of power. What's more, this power is concentrated in very few hands—computational systems make it possible for a few people, or even one person, to exercise power over significant aspects of the lives of billions.

Justifying Power

AI and related technologies are used to exercise power. They have enabled new power relations and intensified other ones. And they allow significant concentration of power. But why does this matter? In particular, what distinctive normative questions are raised by invoking

power? After all, if AI affects people's interests, their choices, their beliefs, and their desires, couldn't we simply evaluate all those effects against, say, a principle of distributive justice? Why not say that we should use AI systems to shape people's lives in ways that, for example, make the worst-off group in society as well-off as they can feasibly be? Or, indeed, why not aim to use AI systems to achieve the goals described in one of the (many) lists of "AI Ethics" principles? Why can't we simply apply one of these standards of *substantive justification* to the use of AI to exercise power?

The answer: the exercise of power by some over others generates presumptive moral objections grounded in individual freedom, social equality, and collective self-determination, which can be answered only if power is used not only for good ends, but legitimately and with proper authority. Space constraints prevent a detailed defence of this thesis; instead, I offer a brief sketch of how such an argument would go.

The first step is to show that the exercise of power generates pro tanto objections, independent of what it is used for. Start with the objection from individual freedom. On a simple, negative conception, one's freedom consists in the absence of external interference in one's choices. A more complex conception would also emphasize the absence of the *risk* of such interference.[35] A further, republican, extension, would add emphasis on the *possibility* of interference (if it is arbitrary).[36] Positive theories of freedom generally add that one must not only avoid interference, but also have an adequate range of options, and the capacities and information necessary to choose wisely among them, based on one's authentic desires.[37] As described in the previous section, AI can be used to directly limit people's negative freedom—to incarcerate and surveil them—as well as to limit their options, and indeed to shape their ability to act authentically to fulfill their desires based on accurate beliefs.

Of course, AI is also used in ways that enhance people's individual freedom—power and freedom are not strict duals. There is a somewhat deeper tension between power and social equality. I understand social equality as the existence of social relations whereby we treat one another as equals.[38] Although in many respects—esteem, affection, and so on—we are *not* equal even in egalitarian societies, in one fundamental and important sense each citizen is the equal of every other. We have the same basic rights and duties, the same standing to invoke the institutions of the state, the same opportunity to participate in them, the same ability to contribute to setting the shared terms of our social existence. The power of some over others involves a hierarchical unidirectional relationship in which A exercises power *over* B, leaving the two in unequal social relations—irrespective of whether A treats B well or poorly. This is a particular concern for the exercise of power by means of AI, both because it relies on expert knowledge that is far beyond the ken of most of those subject to it, and because much of the power exercised by means of AI structures digital environments in which we have long abandoned any pretense of social equality, substituting the aspirationally egalitarian liberal democracies of our offline lives for digital feudalism, subject to the whims of a tiny handful of unaccountable executives in a small district of California.

Social equality can be satisfied if we all have an equal opportunity to shape the shared terms of our social existence, even if we do not actively take up that opportunity. Collective self-determination is to social equality much as positive freedom is to negative freedom: it is about (enough of) us actually positively shaping our world in accordance with our values. If the power of some to shape the shared terms of others' social existence is not an expression

(in some sense) of the collective's will, it is presumptively antithetical to collective self-determination. Our dependence on the whims of those Californian executives undermines our collective self-determination, as well as our social equality.

Even if AI is used to exercise power for goals that serve freedom, equality, and self-determination, or other equally important values, these pro tanto objections focus not strictly on what power is being used to achieve, but rather on the fact that power is being exercised at all. As such, even if AI is used to exercise power for noble ends, these pro tanto objections still apply. They *might* be overridden by the great good being done, but they can be silenced, in my view, only if power is exercised not only for the right ends, but *in the right way*, and *by the right people*. These are the standards of *substantive justification*, *procedural legitimacy*, and *proper authority*.

The standard of substantive justification simply demands that power is used to achieve justified ends. This standard applies whether one invokes power or not—it aims at fair outcomes, the promotion of well-being and autonomy, and the many other goods that we typically aim at in modern liberal democracies.

On the standard of procedural legitimacy, it's not enough to use power to achieve justified ends; it must also be exercised in the right way, by following appropriate procedures. For the exercise of power to be consistent with individual freedom and social equality, it must be subject to strict constraints. We preserve our freedom and ensure that we collectively have power over those who individually have power over us, by limiting their power. We can get some insight into these limits by thinking about the core standards of the rule of law: as well as getting matters substantively right, the governing agent should be consistent, and (morally) treat like cases alike; those subject to the decision should have the opportunity and ability to understand why the decision has been made; standards of due process should be met where feasible, and those exercising power should be subject both to processes of contestation by those subject to their decisions, and to review and potential dismissal on discovery of persistent misconduct.[39] The exercise of power by means of AI has fallen pretty far short of all of these standards; indeed, one might think that when we use complicated ML-based techniques to implement and enforce rules, we are constitutively prevented from meeting these kinds of procedural standards.

On the third, authority standard, it matters not only that power is being exercised for the right purposes and in the right way, but also by the right people. The people exercising power should be those with the authority to do so within that institution. If those who exercise power around here lack authority to do so, then we cannot be collectively self-determining. The criteria for proper authority vary depending on the nature of the institution, but a key, general point is that the more pervasive and important an institution is in the lives of a group of people, the more prima facie important it is that the authority to govern it should stem from them, the people served by that institution. Our informational, material, creative, and other economies are significant parts of our lives. They are also increasingly reliant on digital platforms, which are themselves structured by AI, in particular recommender and search algorithms. And those algorithms are designed and implemented by a small number of employees of private businesses which lack any suitable authority to so extensively shape the shared terms of our social existence. Unauthorized power is a threat to our collective self-determination.

Conclusion

Power over is the social relation where an agent A can significantly affect the interests, options, beliefs and desires of another agent B. AI and related computational systems are being used by some to exercise power over others. They enable new and intensified power relations, and a greater concentration of power. This is especially clear in our online lives, which are increasingly structured and governed by computational systems using some of the most advanced techniques in AI. But it is also apparent in our offline lives, as computational systems using AI are used by powerful actors including states, local government, and employers. Proponents of various principles of "AI Ethics" sometimes imply that the sole normative function of those principles is to ensure that AI is used to achieve socially acceptable goals. They imply that substantive justification is sufficient for all-things-considered justification of these uses of AI. Drawing attention to the ways in which AI systems are used to exercise power demonstrates the inadequacy of this normative analysis. When new and intensified power relations develop, we must attend not only to what power is used for, but also to how and by whom it is used: we must meet standards of procedural legitimacy and proper authority, as well as substantive justification.

Acknowledgments

This chapter has benefited extensively from comments from the editor, Johannes Himmelreich, as well as four anonymous reviewers. Thanks to them all.

Notes

1. E.g. Crawford, Kate. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press; Susskind, Jamie. (2018). *Future politics: Living together in a world transformed by tech*. Oxford University Press; Véliz, Carissa. (2021). *Privacy is power: Why and how you should take back control of your data*. Penguin Books; Nemitz, Paul. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 1–14; Cohen, Julie E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press; Boyd, Ross, & Holton, Robert J. (2018). Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? *Journal of Sociology 54* (3), 331–345; Liu, Hin-Yan. (2018). The power structure of artificial intelligence. *Law, Innovation and Technology 10* (2), 197–229; Bucher, Taina. (2018). *If . . . then: Algorithmic power and politics*. Oxford University Press.
2. Russell, Stuart, & Norvig, Peter (2016). *Artificial intelligence: A modern approach*. Pearson Education.
3. Clarke, Roger. (1988). Information technology and dataveillance. *Commun. ACM 31* (5), 498–512; Nissenbaum, Helen. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2* (1), 25–42; Citron, Danielle Keats. (2008). Technological due process. *Washington University Law Review 85* (6), 1249–1314.

4. E.g., Clegg, Stewart R., & Haugaard, Mark. (2009). *The SAGE handbook of power*. SAGE Publications Ltd.
5. Pansardi, Pamela. (2012). Power to and power over: Two distinct concepts of power? *Journal of Political Power 5* (1), 73–89; Dowding, Keith. (2012). Why should we care about the definition of power? *Journal of Political Power 5* (1), 119–135.
6. Jenkins, Richard. (2009). The ways and means of power: Efficacy and resources. In Stewart R. Clegg & Mark Haugaard (Eds.), *The SAGE handbook of power* (pp. 140–156). SAGE Publications Ltd. For criticism of this kind of view see Young, Iris Marion. (1990). *Justice and the politics of difference*. Princeton University Press.
7. Morriss, Peter. (2002). *Power: A philosophical analysis*. Manchester University Press.
8. Goldman, Alvin I. (1972). Toward a theory of social power. *Philosophical Studies 23* (4), 221–268.
9. Dahl, Robert A. (1957). The concept of power. *Behavioral Science 2* (3), 201–215; Barry, Brian. (1974). The economic approach to the analysis of power and conflict. *Government and Opposition 9* (2), 189–223.
10. For analysis of the debate, see Dowding, Keith. (2008). Agency and structure: Interpreting power relationships. *Journal of Power 1* (1), 21–36, Haslanger, Sally. (2012). *Resisting reality: Social construction and social critique*. Oxford University Press.
11. Haslanger, Sally. (2016). What is a (social) structural explanation? *Philosophical Studies 173* (1), 113–130; Ritchie, Katherine. (2020). Social structures and the ontology of social groups. *Philosophy and Phenomenological Research 100* (2), 402–424. See also the excellent article, "Climate change as a social structural problem" by Max Fedoseev (draft on file with the author).
12. Young, *Justice and the politics of difference*; Haslanger, *Resisting reality*.
13. Dahl, "The concept of power"; Hamilton, Malcolm. (1976). An analysis and typology of social power (part I). *Philosophy of the Social Sciences 6* (4), 289–313.
14. Pettit, Philip. (2008). Dahl's power and republican freedom. *Journal of Power 1* (1), 67–74.
15. Weber, Max (edited and translated by Keith Tribe). (2019). *Economy and society: A new translation*. Harvard University Press.
16. Kernohan, Andrew. (1989). Social power and human agency. *The Journal of Philosophy 86* (12), 712–726.
17. This section draws on a large body of literature describing the social impacts of AI; comprehensive citation would double the length of the chapter. For a very useful description of many of the most important cases, informed by political philosophy, see Susskind, *Future politics*. Other classics of the genre include O'Neil, Cathy. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown; Eubanks, Virginia. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press; Noble, Safiya Umoja. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press; Pasquale, Frank. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
18. Eubanks, *Automating inequality*.
19. Gorwa, Robert, Binns, Reuben, & Katzenbach, Christian. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7* (1), 1–15.
20. Angwin, Julia et al. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, May 23.
21. Brayne, Sarah. (2021). *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press.

22. Susskind, *Future politics*.
23. Brownsword, Roger. (2015). In the year 2061: From law to technological management. *Law, Innovation and Technology 7* (1), 1–51; see also Susskind, *Future politics*.
24. Foucault, Michel. (2010). *The government of self and others: Lectures at the collège de France, 1982–1983*. St Martin's Press.
25. Bucher, Taina. (2013). The friendship assemblage: Investigating programmed sociality on Facebook. *Television & New Media 14* (6), 479–493.
26. Imana, Basileal, Korolova, Aleksandra, & Heidemann, John. (2021). Auditing for discrimination in algorithms delivering job ads. *Proceedings of the Association for Computing Machinery Web Conference*, 3767–3778.
27. Thaler, Richard H., & Sunstein, Cass R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press; Kramer, Adam D. I., Guillory, Jamie E., & Hancock, Jeffrey T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences 111* (24), 8788–8790.
28. Benn, Claire, & Lazar, Seth. (2021). What's wrong with automated influence. *Canadian Journal of Philosophy*, September 24.
29. Yeung, Karen. (2017). "Hypernudge": Big data as regulation by design. *Information, Communication & Society 20* (1), 118–136. In 2022 the field will hold its 17th annual conference: https://persuasivetech.org.
30. On human content moderation: Roberts, Sarah T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press. On information glut: Andrejevic, Mark. (2013). *Infoglut: How too much information is changing the way we think and know*. Routledge. On algorithmic governance: Gorwa, Binns, and Katzenbach, "Algorithmic content moderation."
31. See e.g., Vaidhyanathan, Siva. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press. On the Facebook revelations, see e.g., https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers.
32. Kaptein, Maurits, & Eckles, Dean. (2010). Selecting effective means to any end: Futures and ethics of persuasion profiling. *Proceedings of the Persuasive Technology Conference*, 82–93; Susser, Daniel, Roessler, Beate, & Nissenbaum, Helen. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review 4* , 1–45; Kosinski, Michal, Stillwell, David, & Graepel, Thore. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences 110* (15), 5802–5805.
33. Turow, Joseph. (2011). *The daily you: How the new advertising industry is defining your identity and your worth*. Yale University Press.
34. Benn and Lazar, "What's wrong with automated influence."
35. Kramer, Matthew H. (2008). Liberty and domination. In Cécile Laborde and John Maynor (Eds.), *Republicanism and political theory* (pp. 31–57). Blackwell; Carter, Ian. (2008). How are power and unfreedom related? In Cécile Laborde and John Maynor (Eds.), *Republicanism and political theory* (pp. 58–82). Blackwell.
36. Pettit, Philip. (1997). *Republicanism: A theory of freedom and government*. Clarendon Press.
37. Raz, Joseph. (1986). *The morality of freedom*. Clarendon Press.
38. Anderson, Elizabeth S. (1999). What is the point of equality? *Ethics 109* (2), 287–337; Kolodny, Niko. (2014). Rule over none II: Social equality and justification of democracy. *Philosophy & Public Affairs 42* (4), 287–336.
39. Waldron, Jeremy. (2011). The rule of law and the importance of procedure. *Nomos 50*, 3–31.

CHAPTER 11

AI AND STRUCTURAL INJUSTICE

Foundations for Equity, Values, and Responsibility

JOHANNES HIMMELREICH AND DÉSIRÉE LIM

Introduction

Structural injustice has risen to the top of the agenda. The United States continues to be rocked by police-inflicted violence, especially against people of color; in the background, the American foundational legend—of having discovered a "new world" upon which generations of settlers built their dreams—is falling apart.[1] At the same time, countries in Europe are grappling with disparities in how gender, race, ability, class, and origin affect wages, social recognition, political and economic participation, and casually-inflicted prejudices that perfuse everyday life.[2] Around the globe, the pervasive effects of racism and colonialism have been documented: not only in terms of material differences in outcomes and opportunities, but also in discursive or semiotic differences in how particular groups are portrayed. For example, Black neighborhoods in New York City are advertised as exotic and edgy destinations for the adventurous white tourist (Törnberg & Chiappini, 2020). In short, the Global North is growing the awareness that the *status quo* builds on and manifests various past and present-day injustices and normative deficiencies.

With this growing awareness, students, professionals, academics, and policymakers in the Global North are attending to the role that artificial intelligence (AI) plays in maintaining, entrenching, or even exacerbating this unjust *status quo*.[3] A software feature may be well-intended, and an algorithm considered objective, but when deployed in an unjust *status quo* they will likely perpetuate injustice, or worsen it. AI interacts with unjust social structures—AI exacerbates structural injustice.

What should we make of this idea? The term "structural injustice" has been used to describe a very wide range of different phenomena, including police brutality, unequal health outcomes, pay gaps, and inequalities in educational opportunities. But the theory of structural injustice faces challenges. The theory is epistemically hard. Even if you know there *is*

a problem of structural injustice, you may find understanding what this "problem" *is* just as hard as solving it. Moreover, the theory of structural injustice might seem methodologically deficient. Insofar as it is not couched in terms of agents or institutions, or doesn't specify causal mechanisms, skeptics decry the idea of structural injustice as a set of, perhaps ideologically motivated, causally free-flowing conjectures. The theory of structural injustice is also rhetorically or ideologically disadvantaged. Concerns about structural injustice are often articulated in opposition to a liberal or western mainstream, sometimes as issues of diversity, equity, and inclusion (DEI).[4] But what, exactly, lies behind this language of DEI?

This chapter presents the theory of structural injustice. We explain some of the theoretical foundations of equity and social justice, or, relatedly DEI. We demonstrate that, as opposed to being a mere buzzword, this theory is not a political slogan but a respectable normative and empirical concept. As such, structural injustice ought to inform research and legislation on issues, such as how AI interacts with social identities like gender and race, or how the development of AI reflects existing economic interests. We argue that structural injustice, and AI's capacity to exacerbate it across many spheres of human life, must be taken seriously.

We concentrate here mostly on gender and race. Structural injustices usually attach to salient social categories such as race, gender, age, ability, or sexual orientation. As such, structural injustice has to do with identity. Because race, gender, age, ability, or sexual orientation are central to a person's self-conception or self-image, how one is treated based on these characteristics is a matter of high moral concern and deep emotional valence. But although structural injustice attaches to social identity, this chapter is not concerned with identity politics.[5] Insofar as the chapter aims to make a political contribution at all, ours has a theoretical focus: We are concerned that appeals to the values of DEI or structural injustice in AI are dismissed as a fad, as mysterious methodology, as empty slogans used to virtue-signal, or as sectarian interests pushed by advocacy groups. This is a miskake. We argue that the theory of structural injustice is a useful lens for anyone concerned with AI governance. A theory of structural injustice allows researchers and practitioners to identify, articulate and perhaps even anticipate phenomena and problems that may otherwise go unrecognized. Some basic theoretical ingredients of structural injustice—social-structural explanations and theories of justice—easily extend from gender and race to issues of ability, sexual identity, or economic power.

This chapter covers only limited ground. First, it does not cover business or managerial issues, such as building project teams, nor does it cover the sociology, demographics, or politics of who builds and uses AI today. Making sure that research, design, and engineering teams are diverse is important for addressing structural injustices and for bringing the values of DEI to life. This chapter concentrates instead on the foundations of DEI or social justice: Which theoretical and moral considerations underwrite a concern for DEI? Second, this chapter concentrates on stylized examples. We cannot do justice to the nuanced ways in which structural injustice plays out. This is because social identities are complex, and the experience of structural injustice often attaches to more than one social category at once. We take our examples mainly from a North American context. However, thanks to them being somewhat abstract, the examples should easily travel to other contexts.

In sum, this chapter provides a conceptual lens to bring into sharper focus how AI relates to structural injustice. We contribute to discussions on DEI by conceptualizing present-day calls for diversity, equity, and inclusion in the sphere of AI as *demands of*

justice, rather than a bid to reduce harm or attempts to comply with ethical codes of conduct that are frequently drafted by tech corporations. We do this by giving a primer on the concept of structural injustice as it has been developed by scholars in moral and political philosophy, in particular by Iris Marion Young (Young, 2006, 2011). We argue that this perspective on structural injustice is particularly important for AI. This chapter does not so much highlight the myriad unintended ways in which developing and deploying AI may contribute to structural injustice. Instead, the chapter aims to equip the reader with the theoretical foundations of tools that help to recognize structural injustices as well as with a sense of responsibility to attend to them.

AI and Structural Injustice

As a starting point, we assume that police violence, gender prejudice, and—more broadly—disadvantages because of differences in class, race, ability, or gender are often *structural phenomena*. By this, we mean, first, that they are generally not fully explained by the intentional actions of particular individuals.[6] Police officers may act out of basic instinctive self-preservation or because they internalize problematic informal professional norms, and journalists ask questions that are loaded with gender prejudices because they have reason to believe that viewers care about these questions. Similarly, even the measurable physiological differences—that African American women are 60 percent more likely to have high blood pressure, or that African American children have vastly increased blood lead levels—cannot be linked to any individual act or policy at all. Second, structural phenomena—police violence and elevated lead levels—often seem to have something in common; not only in their outcomes (each concerns a disadvantage on Black people) but also in their causes. A structure is the stipulation of such a general cause.

Well-intentioned individuals—even those who care about the end-goal of racial equality—typically take social structures for granted and accept their constraints and affordances. We tend to work with our world as it is "given." Informal norms, unconsciously internalized expectations, or learned emotional responses are larger than individuals but have a hold on us. Such structural features influence and explain the conduct of individuals and, even more so, aggregate outcomes. Engineers, entrepreneurs, and policy-makers may put great effort behind building products that make the world a better place and that yet still may have the opposite effect. These effects may not be deliberate, but nevertheless, they are profound. To understand this, structural explanations can help. Structural explanations are the central ingredient in a theory of structural injustice (cf. Soon, 2021). Let us turn to an illustrative example.

How structure matters: Example of AI in health care

Hospitals use decision-rules, i.e., algorithms, to identify high-need patients—patients with chronic conditions and complex medical needs—to invite them to special outpatient treatment programs. In 2019, researchers found that an algorithm to identify such high-need patients exhibits "significant racial bias" (Obermeyer et al., 2019). Although more than

46 percent of Black patients should have received the help of such treatment programs, only 17.7 percent did. Black patients were less likely to be selected for the special treatment programs despite there being no difference in underlying health conditions.

The reason for this bias is instructive. On the face of it, from the perspective of researchers and policymakers, the algorithm seems unbiased. In fact, it predicts total medical expenditures without a significant racial difference. The algorithm assigns patients risk scores—this is the prediction it makes—and each of these risk scores is associated with the same average medical expenditures, regardless of whether a patient is white or Black. The algorithm hence correctly predicts, without significant bias, whether or not a patient has high health needs—at least when "high need" is understood as total medical expenditures.

The bias creeps in, however, when we consider the relationship between health and medical expenditures. When developing the algorithm at hand, medical expenditures were falsely taken to be a good proxy for underlying health. On average, however, the healthcare system spends *less* on Black patients than on white patients *at all levels of healthcare needs*. As a result, the algorithm *did* predict *costs* correctly, but given that average healthcare costs are lower for Black patients, the algorithm underestimated the Black patients' healthcare needs. In a way, the algorithm "sees" fewer health needs in Black patients. This lines up with how Black patients' experience their interactions with the medical profession already today (Cuevas et al., 2016). AI thus overlooks Black persons' healthcare needs, rendering them less visible or less urgent.

This is a clear case of racial *injustice* (we leave aside for now what makes it an injustice: why, because not all differences are unjust, is *this* difference an injustice?). This injustice is *structural* because it is best explained with reference to structural features. To get a sense for the power of such structural explanations, consider why, for any given health need, average medical expenditures are lower for Black patients.

One set of hypotheses looks at medical professionals. Doctors need not be outright racists, in the sense of consciously ascribing racial inferiority to Black patients,[7] in order to treat them differently. We inhabit a world where racial categories remain socially salient; that is, our perceived racial difference has a significant impact on how our social interactions unfold across a wide range of contexts. Many people still believe that race is a biobehavioral essence that explains our behavioral and cultural dispositions (Zack, 2014), and they operate on this assumption when dealing with members of other racial groups. Others insist that race is socially constructed, rather than natural or biological. Nonetheless, like other social constructs (e.g., money), race has profound material effects on our lived experience. "Race is," to quote Charles W. Mills, "a contingently deep reality that structures our particular social universe" (1998, p. 48). Given widespread assumptions, stereotypes, prejudices, or generalizations about Black people, including their lifestyles and genetic predispositions, medical staff and professionals might implicitly or unintentionally be less responsive to the needs of Black patients—a problem not only of implicit bias but of discursive norms (Ayala-López, 2018). Thus, even when the locus of the causal mechanism that explains racial differences in medical expenditures is located with individuals—here, the medical professionals—the properties that fuel this mechanism are structural: assumptions, habits of thought, stereotypes, prejudices, discursive norms, or generalizations about Black patients.

Yet the shortcomings of the medical system cannot be solely attributed to the behavior of medical professionals. The second set of hypotheses looks at material features. In many

parts of the United States, white patients are geographically closer to medical resources than Black patients. As Probst et al. (2004) put it, "[d]isadvantage among rural racial/ethnic minorities is a function of place as well as race." This structural feature—geographic location, affordance, and travel costs—can, in part, explain the racial differences in medical expenditures that led to the racial bias in the health AI.

A third set of hypotheses turns to the patients. Black Americans trust the medical system less than other groups (Boulware et al., 2003). This lack of trust can mediate a lack of engagement in care (Eaton et al., 2015), which in turn leads to lower average medical expenditures. The root causes of this lack of trust are largely unclear. Often, researchers seek to explain this lack of trust by pointing to the infamous Tuskegee study. This would be a rather non-structural explanation. However, the Tuskegee study is not sufficiently well-known in the Black American population to directly explain the lack of trust in this way (Brandon et al., 2005). Lack of trust, even if not necessarily a structural feature itself, is likely best explained in reference to structural features—prevailing narratives, expectations, and stereotypes about the medical system. Thus, lack of trust is another structural feature for racial differences in medical expenditures despite identical health conditions.

Despite their variety, these hypotheses still are not exhaustive. Other explanations involve socioeconomic status, gender norms (especially traditional masculine norms that eschew physical vulnerability), lack of awareness of healthcare needs, religious and spiritual attitudes towards medicine, and criminal background (Cheatham et al., 2008). Many of these hypotheses point in the direction of further structural explanations.

This example of AI in healthcare illustrates two key points. First, structural explanations are common in the social sciences. The explanations offered for why Black patients have lower medical expenses are a case in point. How medical professionals respond to Black patients, the geographic distance to hospitals, and the lack of trust in the medical system are components of an underlying social structure. It is such structural features that explain differences in average medical expenditure. Second, not surprisingly then, structural explanations are indispensable when you seek to understand, and anticipate, how AI begets injustice. In the case of the healthcare algorithm, the injustice—not just any old "bias"—crept into production in that the feature of "health" was operationalized as medical expenditures.

The theory of structural injustice

One of the most influential and detailed accounts of structural injustice has been articulated by the philosopher Iris Marion Young. The *structure* of society, as Young characterizes it, is "the confluence of institutional rules and interactive routines, mobilization of resources, as well as physical structures such as buildings and roads" (2006, p. 11). Notice how the structure includes not only rules and conventions—such as implicit assumptions about Black people—but also material properties—the distance to the nearest emergency room. At the same time, *informal social norms and practices* that are not governed by formal rules and conventions, such as stereotypic beliefs, are important constituents of what we mean by "social structure."

Young (2011, pp. 43–45) offers an instructive example of structural injustice in the hypothetical case of a single mother, Sandy. Sandy faces eviction because her apartment building was bought by a property developer. When looking for a new home, Sandy realizes that she

is unable to afford most apartments that are geographically close to her workplace. After some deliberation, Sandy decides to rent an apartment 45 minutes away from her workplace, and reduce the length of her commute by also buying a car that she can drive to work. Unfortunately, Sandy's story has a tragic ending: before Sandy is allowed to move in, her prospective landlord requires that she cough up three months' advance rent. Having spent her savings on the new car, Sandy simply doesn't have the money. She and her children now face the terrifying prospect of homelessness.

Fictional as the case may be, it is no stretch to say that many people have found themselves in similar situations to Sandy's. The case is stylized but sufficiently realistic—perhaps even typical. Two things are noteworthy about this case.

Firstly, the case illustrates how unjust outcomes are brought about by agents with good intentions.[8] The injustice that Sandy endures—being evicted and rendered homeless despite her attempts to make the best of the situation—does not result from the actions and decisions of agents who are out to get her. It is perfectly imaginable that the other characters in this story, such as the landlords, bank employees, mortgage brokers, and property developers, have treated Sandy with decency and respect. They may even be *doing their best* as they face personal struggles of their own. Sandy's original landlord may have decided to sell the building because his financial situation made it impossible for him to maintain it to the standards he should (Young, 2006). The point of Young's story is instead that agents causally contribute to Sandy's homelessness *because* they were acting within the law and according to widely accepted norms and moral rules.[9]

Secondly, identifying any individual action or policy that is wrongful or that causes all that is ethically problematic about this situation seems hard or even impossible. Structural features and not individual actions ultimately explain the predicament that Sandy finds herself in. Sandy's case hence illustrates on an individual level what we have seen in the aggregate in the medical case. As in the case of Sandy, the fact that Black Americans have overall lower medical expenses is not explained by individual choices or policies. Structural injustice stems from many hands and many circumstances. Naturally, material possibilities, stereotypes, geographies, and social norms are not the *only* drivers of racial disadvantage, but they hold much and important explanatory value. The role of agents who are simply "following the rules" or "doing as expected" cannot be overstated.

The case of Sandy hence illustrates the judgment that we started with: No individual acts in a vacuum. The idea of a social structure fills out the situated and context-specific spaces in which we as individuals may often take ourselves to be acting. Accordingly, structural explanations have considerable explanatory power, and are a staple type of explanation in the social sciences (Little, 1991, chapter 5; Haslanger, 2015; Soon, 2021). Structural explanations can explain how Sandy ended up facing homelessness despite no individual wrongdoing and, perhaps even, everyone's best intentions. Structural explanations form the basis for theories of structural injustice.

The concept of *structural injustice*, as we understand it, brings the power of structural explanations to normative analysis. Structural injustice leverages structural explanations and combines them with a theory of justice to enrich analysis, reasoning, and responses to injustice. Three aspects are noteworthy.

First, structural injustice shifts the focus of our normative attention. It starts by seeking to understand injustice instead of theorizing justice. This is a change of focus in both normative as well as empirical theorizing. Normatively, the approach of structural injustice aims to

provide a positive account of injustice; that is, an evaluative theory that conceptualizes injustice not just as the absence of justice or the distance to some ideal of justice. Empirically—our focus in this chapter—structural injustice builds on structural explanations. Structural explanations afford structural injustice great explanatory power. Structural explanations explain why particular nameable groups are persistently disadvantaged. Moreover, structural explanations are unifying. Structural explanations bring into focus commonalities between otherwise disparate-seeming phenomena—the structural features that show up among both causes and consequences. Structural injustice is hence a more holistic or fundamental normative evaluation. This shift in focus—on injustice and on structures—is a hallmark of structural injustice.

Take the example of a police officer who defends herself with objectionable force against a Black man—much like how Amber Guyger fatally shot Botham Jean in his own apartment. Empirically, structural injustice tries to explain how the shooting came about. This violence arises not just from individual behavior but from social norms and practices, and our human tendency to reproduce and reinforce them—by working together and upholding rules and conventions. The social structure constrains and drives individual actions. The structure that underlies police violence constitutes the criminal justice system, the operation of police departments, and negative cultural stereotypes about Black persons. Structural injustice thus identifies these as the targets of reform.

Second, structural injustice is forward-looking. It focuses on structural features. Yet, that does not mean that structural injustice seeks to rectify historic injustice—although it could be complemented in this way. Instead, structural injustice distinguishes between triggering and maintaining causes. The relevant structural causes of racial disadvantage today differ from the past structures that *initially brought about* Black disadvantage. Racial disadvantage may remain alive and well even when state-sanctioned racial discrimination has come to an end (Nuti, 2019). So, in a way, the structural injustice has a temporal dimension and has important historic origins. However, decades away from slavery and the Jim Crow era, racism operates now through novel structural mechanisms that past structures created, such as poverty, new forms of oppression, and different social norms. Structural injustice seeks to understand and reform the structural causes of injustice that operate today.

Finally, structural injustice accumulates. Compared to police violence or Sandy and her family becoming homeless, the example of AI in healthcare appears less troubling. Yet, this impression is misleading. On its own, this disadvantage caused by the healthcare algorithm may seem relatively "small." It may not necessarily make a meaningful difference to health outcomes given that some patients can advocate for themselves and given that there are alternative treatment options. Yet small disadvantages add up over time. Taken together, they explain why the average health of the Black population is lower.[10] Moreover, being less likely to receive preventative healthcare treatment will have negative knock-on effects that lead to further disadvantage. In this way, structural injustice *accumulates* or *compounds* seemingly small disadvantages that derive from agents' tendency to comply with, or operate within the constraints of the status quo.

In sum, the approach of structural injustice has three key features. First, it builds on structural explanations. That is, structural injustice explains phenomena not with reference to individual "micro" attributes (actions of individuals or collective agents) but to broader "macro" attributes (widespread habits of thought, commonplace social practices, and compliance with formal or informal norms). Second, structural injustice is forward-looking. It

aims to identify and explain the maintaining causes of injustices in order to reform them. Third, injustice and disadvantage accumulate. On their own, an individual injustice might be trivial. The significance of structural injustice can be properly appreciated only when looking at the big picture, where individuals must simultaneously contend with *many* types of disadvantage and the constraints they collectively impose. As Young has stated, "The *accumulated effects* [our emphasis] of past decisions and actions have left their mark on the physical world," in a way that forecloses future possibilities" (2011, p. 53).

Structural injustice governs AI

We can now apply the theory of structural injustice to AI. Going beyond the example of AI in healthcare, structural features influence the development and deployment of AI at all steps along the AI pipeline (see the chapter on fairness in this volume for examples). Let us highlight some of these steps.

For starters, structural features influence research agendas, methods, and the choice of problems to tackle. AI research is highly resource-intensive and very expensive. For this reason, powerful economic actors typically decide what problems to tackle and how. Structural features—economic power, political and cultural influence—in part explain which AI is developed and deployed. Looking at who funds and directs AI research institutes that investigate the "ethics" and "fairness" of AI, even at universities that purport to uphold academic freedom and not shy away from critique, you will find it hard to resist the impression that the fox is guarding the hen house (Le Bui & Noble, 2020). Indeed, the fox can just buy the hen house, or, in fact, the whole farm. Next to regulatory capture and cultural capture (Kwak, 2014), in AI governance there is now the problem of academic capture. Similarly, on a smaller scale, individual researchers or public administrators play a causal role in the governance of AI. Again, structural features relating to the social identities of AI's primary movers and shakers—being white, male, having a certain class background—in part explain how AI is developed and used. The structural lens thus brings into focus a strategic analysis of capital interests and ideology, and the causal relevance of social and economic categorical differences between individuals. Combine this structural explanation with a theory of justice, and you may get the result: AI is a form or a tool of structural injustice.

Second, structural injustice is reflected in the data. The case study on AI in health provides an example. Patients who are similarly healthy differ in the medical expenses they incur, depending on their race. Similarly, crime data reflect policing practices just as they reflect actual criminality. In short, social structures explain the patterns of behavior and phenomena that data "represent," and social structures condition practices that generate these data. In the case of Sandy, the available data might fail to account for her plight and the complexity of her case, but Sandy's story is likely to show up in data as in the form of significant disparities between different social groups: geographic segregation by race and class or the intergenerational transmission of wealth and opportunity. In an unjust status quo, data evidences—or can even be a driver of—structural injustice.

Third, social structures shape the understanding and meaning of target variables (Fazelpour & Danks, 2021; Passi & Barocas, 2019). Consider labels such as "gender," "race," "health," "criminality," "creditworthiness," or "academic potential." These target variables

do not merely represent things that are "out there" in the world within a model. Instead, such labels operationalize, encode, or calcify social concepts that are in flux. This is not just a semiotic matter. The possible values of the "gender" variable imply a certain substantive view of what gender is—is it social or biological, a binary or a continuum? Moreover, the meaning of "gender" can significantly change the results of causal analyses (Hu & Kohler-Hausmann, 2020). The analytical relevance and the material effects of the choice of data labels make data labels a matter of structural injustice.

The following two points relate to fairness. There is a consensus in AI governance that AI should be fair. However, this focus on fairness is limiting in important ways. The theory of structural injustice makes clear why pursuing fairness is not enough.

Fourth, structural injustice contributes heavily towards epistemic limits in determining whether an individual was fairly treated. Fairness requires treating like cases alike, but structural injustice makes it hard to tell which cases are alike in the first place. Was Lakisha not hired because of her race, or because Emily was objectively more qualified? If race played a role, then the decision to hire Emily was unfair: Lakisha and Emily are alike (in relevant respects) but were not treated alike. Similarly, did Sandy have to cough up a large deposit for her new apartment because of stereotypes about the financial responsibility of Black mothers? Fairness says that differences should matter to the degree that they exist in a just society. However, in a society rife with structural injustice, it is hard to distinguish between those differences that are caused by injustice and those differences that would persist even if the society were just.[11] This larger epistemic problem for fairness is compounded by a smaller one, namely, the fact that structural injustices accumulate and are therefore hard to track. In sum, structural injustice makes it epistemically hard to be fair (Zimmermann & Lee-Stronach, 2021). Any AI that aims to be fair therefore needs to account for structural injustice. In a slogan, there can be no fairness without an understanding of social structure and what injustice is due to structural maintaining causes. Even as AI might offer new opportunities to formalize and account for structural difference and injustice (Herington, 2020; Kusner & Loftus, 2020), epistemic limitations remain (Ludwig & Mullainathan, 2021).

Fifth, entrenched social structures limit the efficacy of fairness for justice. Fairness often fails to produce justice, similar to how equality fails to produce equity. "Fairness," like "equality" is often understood as a formal condition or an intrinsic virtue of a decision procedure—think of how the maxim to treat like cases alike is a potent source of disparate treatment, under which persons of marginalized social identities are disadvantaged, intentionally or not, because of the failure to recognize salient differences between various groups. In theory, it may be "fair" for prestigious degree programs to only admit students who score high on standardized tests, insofar as "like" candidates are accepted or rejected on the basis of a criterion that applies to all prospective students. All the same, such requirements have had a disparate impact on members of communities who, owing to structural injustices, have lacked the educational resources to score relatively well on standardized tests (rather than being inherently less competent or suitable fits for the university's program). In an unjust status quo, in which injustice is maintained by social structures, a focus on fairness makes it instrumentally hard or perhaps even impossible to promote justice. Justice or "equity," by contrast, may license *unfair* treatment for reasons of justice (Vredenburgh in this volume). Think, here, of preferential hiring and affirmative action. Such measures can affect structural change—by changing stereotypes, enabling role modeling, and affording

recognition—but such measures are arguably unfair in this specific sense. Social structure explains why a focus on fairness in AI might be insufficient for promoting justice.

Finally, social structures affect how AI interacts with the context in which it is deployed (consider cases of disparate impact). The lens of social structural explanations is indispensable for anticipating and analyzing the impact of AI. For example, if automatic license plate readers are deployed among arterial roads where Black Americans are more likely to live, Black Americans are more likely to be subject to the negative consequences of being falsely matched with a license plate on a "hot list" of wanted plates. However, not all effects of structural injustice are so easy to anticipate. Most are not. For example, economic theories of home mortgages hide and entrench structural injustice (Herzog, 2017). If even social scientists struggle to capture structural injustice, project managers, public administrators, or computer scientists cannot hardly be expected to succeed on their own. This is an important governance problem because understanding structural injustice is crucial for anyone seeking to anticipate and analyze the impacts of AI.

Existing Normative AI Governance Frameworks

We have defended the structural injustice approach to the governance of AI. Other approaches are available. The example of medical AI could also be analyzed in terms of harms and benefits—or in terms of values, such as DEI. One could say: The algorithm harmed Black patients. One could also say: The algorithm violated the value of equity, the training data was not really diverse, and the development process not inclusive. Given the existence of such alternative approaches, why choose the approach of structural injustice?

This question is particularly pressing because structural injustice can be hard to grasp—harder than harms and benefits, at the very least. Structural injustice raises formidable methodological challenges for researchers and policy-makers who want to draw on this theory.[12] Given all this, is the approach of structural injustice worth it? We argue that structural injustice is indispensable in the conceptual toolkit. The theory of structural injustice has substantial advantages over other ways of approaching normative and social problems of AI.

Harms and benefits

One framework for analyzing the effects of AI builds on the concepts of harms and benefits. This approach is intuitively compelling and familiar from the ethics of medical research. A particular intervention or artifact—in this case, the use of AI—is evaluated by considering whether there is a risk that individuals would be worse off than they would be without the intervention. Does AI pose threats to their well-being, to their opportunities, or to their health? Are these harms outweighed by benefits?

Such a harms-and-benefits approach surely has its place, but it suffers from severe potential limitations.[13] First, this approach fails to capture ethical problems in full. For starters, it is typically restricted to analyzing *individual* harms and benefits. It may have a harder

time attending to *collective* harms. To go back to an earlier example, marketing historically Black neighborhoods in New York City as dangerous, exciting, and exotic does not seem to harm any individual in particular. In theory, it could benefit a Black property-owner who leases out his apartment on Airbnb. Nevertheless, even language that is not obviously racialized may affect stereotypes and norms about Black Americans, in a way that is not best described as a "harm" let alone one that is separately identifiable and affects specific individuals. Similarly, the use of beauty standards in advertising might not be obviously harmful, let alone be harmful to a *specific* woman, but it may promote distorted beliefs about women as a whole and be a form of structural injustice (Widdows, 2021). Moreover, remember how structural injustice accumulates and compounds. Whereas an approach of harms and benefits identifies individual harms, it may fail to see the fuller picture of how these harms relate. The approach of structural injustice, by contrast, sees disadvantages holistically, compounded by others and adding to other disadvantages in turn.

Second, the concepts of harms and benefits restrict the scope of ethical aspirations and values. Many values are not reducible to harms and benefits.[14] For example, some see it as important that an AI system is explicable or accountable to those subject to it (for clarification on what this means, see other chapters in this section). But the lack of explanations or the absence of accountability does not necessarily constitute a harm. Moreover, some technologies could be beneficial for individuals but still be morally wrong. For example, one can benefit from an intervention to which they *have not* consented. Suppose that, with the help of AI, your employer (or partner) secretly tracks your daily activities, including your dietary and exercise routines. They use these data to serve you lunches that optimize your health and well-being. Despite this benefit, intrusive surveillance without consent is morally off-putting, disturbing, and perhaps morally impermissible. In sum, again a framework of harms and benefits fails to capture the full picture. Governing AI with an eye only to harms and benefits would hence be a mistake. The approach of structural injustice, by contrast, brings into focus structural features such as class interests, economic power, or oppression—concepts that cannot be analyzed purely in terms of harms and their combination.

A third problem with a harms and benefits approach is that a workable account of harms and benefits will need to be accompanied with a theory of how, exactly, harms and benefits ought to be weighed and aggregated. Suppose that some unfortunate individual, Jones, has suffered an accident in the transmission room of a TV station (Scanlon, 1998, p. 235). Jones could be saved from one hour of excruciating pain, but to do so, we would have to cancel the broadcast of a football game, interrupting the pleasure experienced by enthusiastic football fans who are excited about the game. Intuitively, the harm that Jones suffers outweighs the harms that football fans would experience as a result of the canceled transmission. But this judgment depends on a theory of value—and a controversial one. One might argue that, if the number of football fans was sizable enough (e.g., millions of viewers), their collective pleasure might outweigh Jones's suffering. The aggregate benefit is greater than Jones' individual harm. Thus, despite its superficial simplicity, a harms and benefits approach requires a deeper set of principles that describe the aggregation of harms and benefits over individuals and over time. In the case of AI, we need a similar set of principles to justify why certain benefits (e.g., efficiency or economic benefits) ought to be outweighed by considerations of racial justice. But such a set of deeper principles is likely contested and incompatible with the value pluralism—and valuable pluralism—in societies.

The harms and benefits approach is thus neither theoretically simple and, likely, often incompatible with pluralism.

Values and principles

Another approach is that of value statements or ethics principles. Computer scientists, data scientists, or AI practitioners might have to take a pledge on some values or code of conduct, similar to the Hippocratic Oath (e.g., O'Neil, 2016, chapter 11). Organizations might articulate their values, codify them, and bring them to life in their organizational culture and processes. For a while, technology companies, public bodies, and professional associations prolifically listed their principles and values—they may say that technology should be used non-discriminatorily, and that AI should be explicable. *DEI is, in part, an instance of this approach*. An organization may say that they are committed to diversity, equity, and inclusion just as they may say that they are committed to explainable AI—with all the good that this entails: The organization will have processes to determine the meaning of "diversity," "equity," and "explainability" and to make sure its conduct is informed by these values. Moreover, when an organization has publicly committed to such values, it can be held to them, from within as well as from without.

Although such codes of conduct have their place (Davis, 1991; Stevens, 2008), similar to ethical analyses based on harms and benefits, their efficacy is limited (Mittelstadt, 2019; Whittlestone et al., 2019). Statements of values—even if they are articulated in detail, and even if these values are sincerely held and underwritten by a public commitment and organizational structures—are not a viable general approach to AI governance. DEI extends the list of organizational values and principles but suffers from the same shortcomings of this more general approach. There are risks of window-dressing, ethics-washing, and cheap talk. Organizational values might calcify and ethical governance might turn into a compliance exercise of box ticking (Boddington, 2020). More importantly, statements of values or principles are not a viable ethical framework for the governance of AI for three reasons.

First, some organizations may see AI ethics as their mission but cannot, on their own, bring values to life. Examples here are professional organizations, such as the Association of Computing Machinery (ACM), the American Statistical Association, or the American Society for Public Administration. Such organizations lack the processes to explicate, role-model, incentivize, or enforce the values that they give themselves—processes that are crucial for accountability and for codes to be effective in a governance context (Gasser & Schmitt, 2020; Stevens, 2008).

Second, the task of governance may involve many actors with diverging or competing values. The approach of articulating ethical values and principles, and bringing them to life, cannot do much to reconcile differences. Similar to the problem of the harms and benefits approach, this is a major shortcoming because value pluralism is central to many issues—from the ethics of autonomous vehicles to the value alignment problem (Gabriel, 2020; Himmelreich, 2020).

Of course, such pluralism could be accommodated, if the principles are limited to some basic consensus. Such consensus or minimal principles could be ethical guardrails or democratic principles and values. In this vein, the state—and especially its courts and executive

agencies—often purport to be based on values of this sort, neutral values or consensus values. But whether such neutrality is feasible and whether it is desirable is questionable (Wall, 2021, sec. 3.1). Moreover, there is likely a tradeoff between a set of values that is neutral and that can find consensus on the one hand, and a set of values that is interesting, promotes justice, and guides actions. Principles and values that may pass muster and count as neutral enough—such as respect for value pluralism or freedom of speech—might just not be informative enough to guide actions or regulations, let alone promote justice (Himmelreich, 2022).

Justice

A third possible evaluative framework centers around the idea of *justice*. Theories of justice broaden our understanding of the *unit of evaluation*—beyond harms and benefits—and they account for the *foundations of values*—beyond merely listing or stating them. Theories of justice aim to orient reasoning and discussions of regulation and policy—and in this sense inform policy-making—while they aim, at the same time, to have a broad appeal. This is because theories of justice are meant to regulate issues of *common* concern. The idea is to have a theory to regulate how we get along, in a way that makes room for conflicting views and beliefs about what is morally right and good. Theories of justice start from the idea that there are moral conflicts and deep disagreements on matters of common concern. All this makes theories of justice the most useful and fitting lens through which to analyze normative issues of AI.

In both ordinary and philosophical discourse, the words "ethics" and "justice" are frequently run together. It is commonly assumed that what is *ethical* is also *just*: the measure of justice, in a particular society, is how ethically the state treats its citizens and other persons who are subject to its power. Politics, then, is seen as amounting to applied ethics. Pressing political questions—e.g., whether capital punishment is morally permissible, or whether a state may block immigrants from entry—are ethical questions similar to, say, whether eating meat or buying fast fashion is morally permissible.

Justice, by contrast, is concerned with normative requirements or considerations that are different—or have different emphasis—from those of ethics more broadly. As Bernard Williams writes, political philosophy should "use distinctively political concepts, such as power, and its normative relative, legitimation" (2005, p. 77).[15] According to Williams, the structures we live within must *make sense* to us, to the extent that we are able to see why it would be wrong for us to reject or resist those structures. For theorists like Williams who strongly distinguish between "ethics" and "politics," the central puzzle of politics is not "are we being treated ethically?" but rather, "can we *make sense* of the exercises of power that we are routinely subject to?" Justice, therefore, is not a matter of compliance, of applying principles, or of living by values but it is instead decidedly practical and involves processes, such as public deliberation and contestation. Put differently, justice is *not* primarily about treating people in line with *ethical* principles: instead, for justice, exercises of power must be *justified* to those subject to them. The approach of justice brings into focus questions of who may issue rules and whose wordcounts (authority), the processes in which such rules are made and enforced (legitimacy), and the reasons for the rules, decisions or actions (justification).

The task of theories of justice is to normatively ground the regulation and interrogation of power (see the respective chapter on power in this volume). This applies to the legal system as well as to the markets and the economic system. Such systems cannot take *any* form that power-holders desire. They must be built or maintained in a way that *makes sense* to the persons who live within them. This is because, as Rawls insists, power must be legitimated to us because of its "profound and present" effects on our lives—our life-prospects, goals, attitudes, relationships, and characters (Rawls, 1971, Sect. 2). Suffice to say, a society that persistently disadvantaged or subordinated persons on the basis of gender or race would be extremely difficult, if not entirely impossible, to legitimate. It is more than reasonable, for a Black man who is disproportionately subject to state-sanctioned police violence, to ask *why* he should be required to accept these social conditions. Quite obviously, a racist social arrangement would not "make sense" to him or others in his position. Either way, then, theories of justice—whether as an extension or a reform of existing theories—are the right kind of theoretical framework to address fundamental normative issues in the governance of AI.

Theories of justice spell out ideas of equality, freedom, and community. Such theories explain why racism is wrong, why colonialism is wrong, what oppression is, how it can be overcome, what an absence of these wrongs would look like, and why such a state would be desirable, even under conditions of pluralism. Theories of justice also delineate and ground liberties. They explain, for example, when and why citizens have a right to an explanation from courts or administrative bodies.

Therefore, the approach of justice has several advantages over the alternative approaches. In contrast to harms and benefits, it broadens the unit of evaluation. In contrast to the approach of listing values and principles, the justice-based approach aims not at values themselves but the underlying reasons for values—it answers the question of why we need values like explainability, accountability, or equity at all. And in contrast to both alternative approaches, justice puts dilemmas and conflicts *between* individuals front and center. The approach of justice aims to regulate matters of common concern. It presupposes and respects a meaningful degree of conflict and disagreement.

Diversity, equity, and inclusion in a theory of justice

The approach of justice hence is more foundational than the alternative two approaches. We hope it can also be useful. To illustrate, consider how the approach of structural injustice in particular recognizes calls for greater DEI as demands of justice. Articulated as demands of justice, DEI is not an attempt on the part of marginalized social groups to secure more power, resources, and advantage for themselves, as it is often uncharitably interpreted. Nor is it a mere matter of generosity or beneficence that would help to make the world a morally better place. From the perspective of justice, the values of DEI stand on reasons that should have a hold on everyone regardless of their self-interest. Such reasons of justice weigh more heavily than the reasons to help others in need. Moreover, reasons of justice are important because the institutions in an unjust society often lack legitimacy and authority.

To illustrate how the demands of DEI can be seen as demands of justice, we need to unpack the content of DEI in more detail. In our view, "diversity," "equity," and "inclusion" are separate but closely interrelated ideas. Out of the three, we understand equity as

the fundamental one. Not surprisingly, egalitarian theories of justice require (or assume), among other things, some form of gender and racial equity for justice to obtain. Women and persons of color cannot be asked to accept a society that persistently subjects them to disadvantage. In this sense, in trying to make society acceptable to everyone, DEI—and especially equity—formulates a partial ideal of justice.

Moreover, the values of diversity and inclusion are instrumental for, or even constitutive of, achieving this ideal of equity. For example, the tech industry remains heavily male-dominated, and women have been twice as likely to leave as men.[16] Arguably, gender equity can only be achieved if tech corporations aim for gender diversity—for example, by actively recruiting and retaining women. In the absence of women practitioners who are attentive to existing gender-based disparities, there is considerable risk that AI will inadvertently perpetuate structural gender injustice. At the same time, to be truly inclusive towards women, tech corporations must reconsider their professional norms and practices, and how those may be hostile or exclusionary to women. Sexist stereotypes about women's inherent lack of suitability for science, technology, engineering, and math (STEM) fields, which can lead to biased and unfair treatment towards women employees, must also be resisted.[17]

Approaching AI ethics through the lens of justice may not just vindicate and give structure to the ideas of DEI. The approach of structural injustice may also enrich and clarify their content. The idea of "diversity" can be—and for the purposes of American constitutional law, often has been—understood as symmetric: A relatively homogenous group of students is technically made more "diverse" by anyone who differs from the group members in one of many possible dimensions, but the justice-based approach will interrogate the reasons for valuing "diversity" and, from there, inform our sense of *which* diversities are relevant—this is why gender and racial diversity have taken center stage in attempts to "diversify" particular spaces, rather than diversity in properties like hair or eye color. As we have seen, justice takes into account persistent, and even historic, disadvantages that are connected to these social identities—see, for example, Charles W. Mills' reminder of the importance of corrective justice (2017, p. 208). To make matters worse, "diversity" is often understood as an individualistic notion. An individual improves a group's diversity simply in virtue of their intrinsic properties. On this view, add an engineer of color to the team, and the work of DEI is done. By contrast, an approach of justice—especially an approach of structural injustice, such as that of Iris Marion Young (2009), looks to overall social conditions that determine or constrain our possibilities, not merely to individual contributions to individual groups. It serves as a reminder that structural disparities often continue to obtain even when certain individuals from diverse backgrounds may achieve great success within their occupation. Existing theories of justice and structural injustice hence may not only ground the ideas of DEI, but may also enrich and clarify their content.

Conclusion

This chapter defends two important parts of a framework for the governance of AI: structure and justice. We argued, first, that an approach to the governance of AI should avail itself of the analytical benefits of structural explanations, and, second, that the evaluative component of such a framework should be provided by a theory of justice. We illustrated

the advantages of structural explanations and how an approach of structural injustice recognizes and advances the values of DEI.

In conclusion, it is time to look forward. First, we sketch a relevant theoretical question in how theories of justice relate to AI. Second, in a more practical vein, we outline how the approach of structural injustice can be accompanied by an understanding of responsibility.

AI as part of the structure?

A theory of AI justice can take two forms. First, existing theories of justice can be applied to AI. This strategy has clear prospects. Theories of justice can explain why AI should be explainable just like other decisions should be explainable. The demand for explainability in AI may fall out of a more general obligation to explain judicial and administrative decisions to those subject to them.

Second, instead of merely applying theories of justice to AI, AI can be the novel subject of a theory of justice. This second theoretical avenue is motivated by the thought that, broadly put, AI changes the nature of society—the social structure, the subject of justice. AI not only augments existing social practices, and AI does not just realize existing social functions in a novel technological way. Instead, the idea is that AI raises questions of justice just like the tax system, the financial system, the criminal justice system, the judicial system, or the social system—e.g., interpersonal interactions and family—raise substantive questions of justice. On this second approach, AI is seen as part of the social structure.

Justice and structure are closely related. A focus on justice requires a focus on structure. This idea goes back to John Rawls whose work is foundational for contemporary theories of justice.[18] In Rawls' picture, the subject of justice is the *basic structure* of a society; that is, all institutions, taken together and over time—including political, legal, economic, social, and civil systems—that form the backdrop for life in this society.

Rawls—not surprisingly—assumes a narrow and dated idea of the basic structure in several ways. First, Rawls pays relatively little attention to the individuals in the structure (Cohen, 1997). Individuals need to obey the law but they are not subject to demands of justice directly. Rawls concentrates instead largely on "the political constitution and principal economic and social arrangements." However, Rawls's articulation of the basic structure falls short in a different respect. While it briefly includes the "monogamous family" as an example of a "major social institution" (Rawls, 1971), the feminist philosopher Susan Okin sharply critiques Rawls's inattention to families, including family structures that are not necessarily "monogamous" (Okin, 1989, p. 93) and provides a detailed elaboration on how family units can be major sites of social inequalities. Women are typically assumed to be responsible for performing the lion's share of domestic chores and caregiving labor, which may severely hamper their opportunities relative to men. Additionally, families are the "first schools of moral development, where we first learn to develop a sense of justice" (Okin, 1989, p. 31). Here, the implication is that unjust family structures can hamper or distort our nascent capacities to observe just terms of cooperation under circumstances of conflicting interests and scarce resources. For these reasons, Okin believes that families must also be regulated by principles of justice. It is best to think of the basic structure as including not only the legal system and the economy, but also the family. The idea of the basic structure is

thus not only closely related to ideas of justice, what counts as the basic structure is also not set in stone, not even when we think about justice in the tradition of Rawls.

We take Rawls's willingness to adapt his definition of the "basic structure," in the face of societal upheaval, to be a sign that the idea of justice can and should be responsive to the growing ubiquity of AI.

In the age of AI, and the overwhelming power of "Big Tech" to shape our everyday lives, AI might have to be seen as an important component of this "basic structure." The question is thus: Should AI be seen as part of the basic structure?

Iason Gabriel (2022) argues that it should. He argues that the basic structure of society is best understood as "a composite of *socio-technical systems*" that bring about "new forms of stable institutional practice and behavior." It would be a mistake to think of "the political constitution and principal economic and social arrangements" as somehow removed or insulated from its interactions with technology. Instead, AI "increasingly *shapes* elements of the basic structure in relevant ways." It *mediates* the distribution of basic rights and duties, along with the advantages of social cooperation (Gabriel, 2022). The healthcare algorithm above is only one example. Public services—from policing to welfare, housing, and infrastructure—are increasingly automated, with profound aggregate effects on those who are already disadvantaged (Eubanks, 2018).

However, the question of how to best develop theories justice for AI—by applying existing theories or by rethinking conceptions of the basic structure—is still open. It is a crucial question for ongoing normative and empirical research.

AI justice via responsibility

In closing, we want to highlight one practical upshot of structural injustice. Structural injustice emphasizes individual responsibility (Goodin & Barry, 2021; McKeown, 2021). The cultivation of practices of responsibility is thus an important topic for AI governance.

Injustice raises a question of responsibility: Who is responsible for rectifying structural injustice? Two answers may immediately come to mind.

The first answer is that *everyone* is responsible *together*. Institutionally, this means that the state might be responsible. States are already responsible for discharging duties of justice to their citizens. States implement ideas of distributive justice: they tax, subsidize, and incentivize to allocate resources and opportunities. On this view, private entities, by contrast, are not bound by considerations of justice. Thus, they are not responsible for rectifying structural injustice.

The second answer is that *nobody* is responsible. After all, structural injustice emphasizes that agents merely contribute to but do not cause structural injustice. In this way, individuals are not wronging anyone. It may thus be inappropriate to hold individuals responsible, let alone blame them for structural injustice.

Both these answers are unsatisfactory.[19] That states alone are responsible for rectifying structural injustice is unconvincing. Our chapter has emphasized the growing power and profound effects that AI has on individuals' life-prospects. A state-centric view of political responsibility would betray an overly limited view of the social structure. Many non-state actors—"Big Tech" and billionaires—shape AI, whether AI itself is part of the basic structure or not.

Moreover, both answers assume an overly narrow view of responsibility. They falsely assume that responsibility entails blameworthiness. Instead, we can distinguish between *attributive* and *substantive responsibility*. Attributive responsibility "helps us decide to whom we should attribute (retrospectively) praise or blame for a particular state of affairs" (Parrish, 2009). Substantive responsibility, by contrast, refers to what people are required to do (Scanlon, 1998, chapter 6).

Suppose that, in a fit of anger, I spitefully pour water onto your laptop to damage it. Here, I am clearly *attributively responsible* for destroying your laptop; it is appropriate for you to blame me for what I have done. Of course, I am also *substantively responsible* for paying for your laptop. By contrast, suppose instead that I have accidentally spilled water on your laptop. It was an accident with no fault on my part. In this scenario, I am *not* attributively responsible for damaging your laptop but I can still be substantively responsible, due to the causal role I have played in damaging the laptop. Paying for the repair is still on me (on the responsibility for AI and the different roles of responsibility see Himmelreich and Köhler forthcoming).

For structural injustice, it seems best to focus on substantive responsibility. Even if individuals are not to blame for structural injustice, they can still be responsible for rectifying it. The goal of responsibility here is to identify and address individual actions, which may be blameless, but that may have generated injustice. Moreover, we can ask what social changes will prevent—or at least reduce the likelihood of—future structural disadvantage, even if they did not causally contribute to existing structural injustice. These are two paths of assigning responsibility, without blame, that operationalize the theory of structural injustice to affect structural reform.[20]

In conclusion, the overall picture that we hoped to sketch in this chapter is this: Structural injustice offers an analytical framework and a normative framework—via structural explanations and a theory of justice respectively. Both of these are, or so we have argued, indispensable for a normative theory in the governance of AI. Moreover, structural injustice offers a useful practical approach by aiming to identify responsibilities of individuals to reform where they can and to rectify without blame.

Notes

1. This legend may never have been plausible outside the lived experience of a middle- to upper-class white population.
2. Such "everyday" instantiations of prejudice are popularly known as "microaggressions." For a detailed philosophical account of microaggressions and the injustice they may perpetuate, see Rini (2020).
3. We understand "AI" as statistical methods that use machine learning (ML) to make data-based predictions or decisions. With the increasing availability of data, and the decreasing costs of storage and computing, the possible uses of AI have dramatically increased.
4. Our argument extends to the value of accessibility. The use of "DEI" instead of "DEIA" is not intended to make a relevant semantic or pragmatic difference. For the purposes here, we understand the "A" to be entailed by, or synonymous to, the "I".
5. Perhaps structural injustice should *not* be approached as a matter of identity at all, but as a matter of difference (Young, 2009).

6. Intentional racism etc. is a problem—and a glaring one. What is wrong here is relatively easy to explain, in contrast to structural injustice.
7. See Appiah (1990) for an influential analysis of racism. Roughly speaking, Appiah defines "intrinsic racism" as the belief that a racial group is intrinsically superior or inferior to others.
8. Again, we take for granted here the evaluation that would be given by a theory of justice—that Sandy's situation is an injustice.
9. The formulation here should draw attention to the methodological challenges of structural explanations. Although individual agents causally contribute to the outcome or phenomenon, the structural features are an essential part of the explanation, nonetheless.
10. The causal story is, of course, more complex. Health outcomes depend not only on material and economic factors, but also on social factors.
11. Audit studies are one way of identifying such differences. The "Lakisha" in this paragraph is a reference to one prominent audit study (Bertrand & Mullainathan, 2004). However, their methodology is controversial.
12. Here, we refer to a challenge that we cannot pursue in this chapter. It divides up into two sets of issues. First, what is the relevant knowledge required to grasp structural injustice? Could the knowledge be conveyed by social-scientific theories? If so, which ones? Alternatively, is such knowledge impossible to quantify or formalize? Second, and relatedly, how can this knowledge be obtained? Do some individuals, because of their social position or role, have better access to the relevant knowledge?
13. A good example of a harms-based framework in AI that avoids some of these problems is Microsoft's Azure Architecture Application Guide (Microsoft, 2021).
14. Here, we understand "harm" as the setting back of one's interests.
15. We use Williams here as an illustrative slogan because we disagree with the metanormative view—political realism—that this quote conveys.
16. See https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/.
17. The pursuit of diversity and inclusion also relates to other ideas and values, not just to justice. Although we focus on diversity and inclusion as requirements of justice, there can be other valuable instrumental benefits to diversity and inclusion, such as greater epistemic performance of a group. For example, with more diverse input, tech corporations might get better at developing helpful algorithms.
18. Rawls asserts, "the primary subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the advantages from social cooperation" (1971). While we do not seek to defend a Rawlsian theory of justice, nor is it necessary to spell out the exact principles of justice that Rawls has proposed, his work is foundational to how we understand the point and purpose of a theory of justice.
19. Given the purposes and limitations of this chapter, our treatment of the relevant substance here will be very superficial.
20. To be clear, this does not mean that everyone is equally responsible for rectifying structural injustice. Depending on their social position, some participants in the relevant structure may have made greater causal contributions than others, and consequently have more burdensome responsibilities. Some participants may also have more power than others to rectify structural injustice. Finally, some agents may also still have attributive responsibility and be blameworthy for the injustices in which they are involved.

References

Appiah, K. A. (1990). Racisms. In D. T. Goldberg (Ed.), *Anatomy of racism* (pp. 3–17). University of Minnesota Press.

Ayala-López, S. (2018). A structural explanation of injustice in conversations: It's about norms. *Pacific Philosophical Quarterly 99*(4), 726–748. https://doi.org/10.1111/papq.12244.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review 94*(4), 991–1013.

Boddington, P. (2020). Normative modes: Codes and standards. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 123–140). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.7.

Boulware, L. E., Cooper, L. A., Ratner, L. E., LaVeist, T. A., & Powe, N. R. (2003). Race and trust in the health care system. *Public Health Reports 118*(4), 358–365. https://doi.org/10.1093/phr/118.4.358.

Brandon, D. T., Isaac, L. A., & LaVeist, T. A. (2005). The legacy of Tuskegee and trust in medical care: Is Tuskegee responsible for race differences in mistrust of medical care? *Journal of the National Medical Association 97*(7), 951–956.

Cheatham, C. T., Barksdale, D. J., & Rodgers, S. G. (2008). Barriers to health care and health-seeking behaviors faced by Black men. *Journal of the American Academy of Nurse Practitioners 20*(11), 555–562. https://doi.org/10.1111/j.1745-7599.2008.00359.x.

Cohen, G. A. (1997). Where the action is: On the site of distributive justice. *Philosophy & Public Affairs 26*(1), 3–30.

Cuevas, A. G., O'Brien, K., & Saha, S. (2016). African American experiences in healthcare: "I always feel like I'm getting skipped over." *Health Psychology 35*(9), 987–995. https://doi.org/10.1037/hea0000368.

Davis, M. (1991). Thinking like an engineer: The place of a code of ethics in the practice of a profession. *Philosophy & Public Affairs 20*(2), 150–167.

Eaton, L. A., Driffin, D. D., Kegler, C., Smith, H., Conway-Washington, C., White, D., & Cherry, C. (2015). The role of stigma and medical mistrust in the routine health care engagement of black men who have sex with men. *American Journal of Public Health 105*(2), e75–e82. https://doi.org/10.2105/AJPH.2014.302322.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass 16*(8), e12760. https://doi.org/10.1111/phc3.12760.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Gabriel, I. (2022). Towards a theory of justice for artificial intelligence. *Dædalus 151*(2), 218–231.

Gasser, U., & Schmitt, C. (2020). The role of professional norms in the governance of artificial intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 140–159). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.8.

Goodin, R. E., & Barry, C. (2021). Responsibility for structural injustice: A third thought. *Politics, Philosophy & Economics 20*(4), 339–356. https://doi.org/10.1177/1470594X211027257.

Haslanger, S. (2015). What is a (social) structural explanation? *Philosophical Studies 173*(1), 113–130. https://doi.org/10.1007/s11098-014-0434-5.

Herington, J. (2020). Measuring fairness in an unfair world. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 286–292. https://doi.org/10.1145/3375627.3375854.

Herzog, L. (2017). What could be wrong with a mortgage? Private debt markets from a perspective of structural injustice. *Journal of Political Philosophy 25*(4), 411–434. https://doi.org/10.1111/jopp.12107.

Himmelreich, J. (2020). Ethics of technology needs more political philosophy. *Communications of the ACM 63*(1), 33–35. https://doi.org/10.1145/3339905.

Himmelreich, J. (2022). Against "Democratizing AI." *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01357-z.

Himmelreich, J., & Köhler, S. (forthcoming). Responsible AI through conceptual engineering. *Philosophy & Technology*. https://doi.org/10.1007/s13347-022-00542-2.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 513. https://doi.org/10.1145/3351095.3375674.

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature 578*(7793), 34–36. https://doi.org/10.1038/d41586-020-00274-3.

Kwak, J. (2014). Cultural capture and the financial crisis. In D. Carpenter & D. A. Moss (Eds.), *Preventing regulatory capture: Special interest influence and how to limit it* (pp. 79–81). Cambridge University Press.

Le Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence: On the limits, failings, and ethics of fairness. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 161–179). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.9.

Little, D. (1991). *Varieties of social explanation: An introduction to the philosophy of social science*. Westview Press.

Ludwig, J., & Mullainathan, S. (2021). Fragile algorithms and fallible decision-makers: Lessons from the justice system. *Journal of Economic Perspectives 35*(4), 71–96. https://doi.org/10.1257/jep.35.4.71.

McKeown, M. (2021). Structural injustice. *Philosophy Compass 16*(7), e12757. https://doi.org/10.1111/phc3.12757.

Microsoft. (2021, November 12). Types of harm—Azure application architecture guide. Redmond, WA. https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm.

Mills, C. W. (1998). *Blackness visible: Essays on philosophy and race*. Cornell University Press.

Mills, C. W. (2017). *Black rights/white wrongs: The critique of racial liberalism*. Oxford University Press.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Nuti, A. (2019). *Injustice and the reproduction of history: Structural inequalities, gender and redress*. Cambridge University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

Okin, S. M. (1989). *Justice, gender and the family*. Basic Books.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Parrish, J. M. (2009). Collective responsibility and the state. *International Theory 1*(1), 119–154. https://doi.org/10.1017/S1752971909000013.

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. https://doi.org/10.1145/3287560.3287567.

Probst, J. C., Moore, C. G., Glover, S. H., & Samuels, M. E. (2004). Person and place: The compounding effects of race/ethnicity and rurality on health. *American Journal of Public Health 94*(10), 1695–1703.

Rawls, J. (1971). *A theory of justice*. 1999 revised edition. Harvard University Press.

Rini, R. (2020). *The ethics of microaggression*. Routledge.

Scanlon, T. (1998). *What we owe to each other*. Belknap Press of Harvard University Press.

Soon, V. (2021). Social structural explanation. *Philosophy Compass 16*(10), e12782. https://doi.org/10.1111/phc3.12782.

Stevens, B. (2008). Corporate ethical codes: Effective instruments for influencing behavior. *Journal of Business Ethics 78*(4), 601–609. https://doi.org/10.1007/s10551-007-9370-z.

Törnberg, P., & Chiappini, L. (2020). Selling black places on Airbnb: Colonial discourse and the marketing of black communities in New York City. Environment and Planning A: *Economy and Space 52*(3), 553–572. https://doi.org/10.1177/0308518X19886321.

Wall, S. (2021). Perfectionism in moral and political philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Fall. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/perfectionism-moral/.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* January, 195–200. https://doi.org/10.1145/3306618.3314289.

Widdows, H. (2021). Structural injustice and the requirements of beauty. *Journal of Social Philosophy 52*(2), 251–269. https://doi.org/10.1111/josp.12369.

Williams, B. (2005). *In the beginning was the deed: Realism and moralism in political argument*. Princeton University Press.

Young, I. M. (2006). Responsibility and global justice: A social connection model. *Social Philosophy and Policy 23*(1), 102–130. https://doi.org/10.1017/S0265052506060043.

Young, I. M. (2009). Structural injustice and the politics of difference. In T. Christiano & J. Christman (Eds.), *Contemporary Debates in Political Philosophy* (pp. 362–383). John Wiley & Sons. http://onlinelibrary.wiley.com/book/10.1002/9781444310399.

Young, I. M. (2011). *Responsibility for justice*. Oxford University Press.

Zack, N. (2014). The philosophical roots of racial essentialism and its legacy. *Confluence: Journal of World Philosophies*, *1*, 85–98. https://scholarworks.iu.edu/iupjournals/index.php/confluence/article/view/522.

Zimmermann, A., & Lee-Stronach, C. (2021). Proceed with caution. *Canadian Journal of Philosophy*, 1–20. https://doi.org/10.1017/can.2021.17

CHAPTER 12

BEYOND JUSTICE

Artificial Intelligence and the Value of Community

JURI VIEHOFF

Introduction

When an algorithm-powered software unfairly rejects your loan application because you live in a poor neighborhood, or denies bail because of your race, or violates your privacy by illicitly collecting and aggregating data from multiple sources, then it is *you* who is harmed, and your rights that are violated. Many of the ethical challenges that the adoption of artificial intelligence (AI) has created can be described in terms of such individual harms and rights. It is to such instances that the language of justice applies, if not exclusively, then at least most comfortably: it is a matter of justice that our rights are protected, and it is the job of a theory of justice to tell us what rights and duties govern interpersonal relations and what political institutions must be like to protect our rights.

But not everything we care about in political life is easily described in terms of individual rights. Consider the revolutionary triad of "Liberty, Equality, Fraternity": much recent political philosophy is an attempt to reconcile the values of liberty and equality through an account of equal rights and entitlements (Rawls, 1999; Dworkin, 2000). But the third ideal, *fraternity* (or "political community"[1]), seems different. Community is the property of a group rather than of any individual. It is *beyond* justice—not in the sense that it tracks less important interests, but in the sense that in understanding the nature of the ideal we must look beyond individual rights. Political community is an ideal worth investigating: if you think there can be things that make communities better or worse, and if you think it is valuable to have a good community, then we should figure out what exactly it is that makes a community "good" so that we know what exactly our policies should aim for.

AI will lead to significant societal transformations, many still uncertain. But we can speculate on what these will be. Such an exercise is productive if it allows us to formulate strategies to protect valuable practices and mitigate negative effects. This article is such an instance of productive speculation: it assesses how AI technology will shape our ability to realize

community in contemporary polities. To do so, we first need an account of community. We will then assess how AI shapes community; we begin by sketching some positive effects that AI could have. We then contrast this hopeful vision with AI's challenges by exploring how its implementation has threatened community in existing, imperfect polities through *entrenchment* and *dispersal* of anti-community pathologies and *decommunitarization*. Next, we will look at possible structural shifts, namely "datafication," "automation," and a "disappearing public sphere" and what these may entail for the ideal's realization. Finally, the chapter will summarize the argument and points to some policies to both mitigate AI's threats and harness its possibilities.

There are three main takeaways. First, community constitutes an independent dimension of value in political life that cannot be captured completely by the demands of justice. Consequently, second, our assessment of AI's impact should not be limited to questions of individual rights and personal goods like autonomy, privacy and non-discrimination. We must, in the case of large polities, also investigate AI's impact on core elements of community (e.g., informal social norms about equal standing, communal attitudes about reciprocity) and practices of joint deliberation. Third, in order to address the already exiting (mostly negative) effects of AI on community, we must rely on both received strategies and develop new ones. For some consequences of AI, renewed focus on standard welfare policies and genuine democratic deliberation may be enough. But where AI creates new forms of social power and control that affect community, we may need to develop and implement genuinely new forms of collective political action to control where and how AI is implemented.

Community as an Ideal

Most people have an intuitive grasp of when community is present in small groups. Think, for example, of somebody describing their rural place of upbringing: "It isn't just a village, it is a *community*!" In a village community, people share a sense of unity or belonging together: they don't just live side by side. Community is a matter of, first, members' self-understanding and beliefs and, second, their actual behavior toward each other (Tönnies, 2002; Weber, 2002).

On the first: members believe that there exists some *shared feature* amongst them and from this they derive a belief in shared fate or *identification*. Second, they have *special concern*, such as a motivation to assist other members (Shelby, 2005, p. 63). Third, *caring*: members experience both positive and negative emotions on behalf of other members and the group (G. A. Cohen, 2009, p. 34). Fourth, *agency*: members see themselves as jointly having agency. As a group, they can have aims and goals, can act, and can be acted upon (Zhao, 2019, p. 54). Fifth, community contains a commitment to a type of equal relationship in that members reject differences in social status and privilege (D. Miller, 1989, p. 230). Finally, members *value* holding these beliefs and having these feelings: they are not alien to them, but positively endorsed and believed to be held for reasons.

On the second: Our villagers will have an actual community only if members act on these feelings and beliefs. Thus, *identification*, *special concern*, and *care* exert pressure in the

direction of communal sharing and wide reciprocity (Fiske, 1993, p. 693ff.): when a house in the village community is hit by lightening, everybody chips in to rebuild it. *Agency* and *equality of status* shapes norms of joint deliberation and decision-making: when villagers need to decide on whether to build a new road, they choose together and everybody gets a say in the townhall meeting. Beyond this, community prevails when the attitudes of each member integrate with, and sustain, those of others: villagers know that her neighbors hold these attitudes, and they know that others know that they and others hold them, and, moreover, they hold them in part *because* others hold them. Thus, community prevails when there is *mutual* identification, *mutual* concern, *mutual* care, and so on.

When and why is community valuable? First, community is instrumentally valuable in that it typically creates the social preconditions for the provision of public goods, ranging from environmental protection to public infrastructure. Public goods are non-excludable and/or require cooperation, which are easier to realize if community is present. Second, there is also a non-instrumental value to community: it is good to share such bonds with others, to care for and be cared for by those with whom we share a social world. Political community of the right kind gives us a sense of purpose and belonging, and it allows members to pursue shared options that are valuable because they are shared (Margalit & Raz, 1990).

Now village communities are small and intimate. Is it possible to extend this ideal to complex, large-scale societies with coercive state institutions? The answer is "yes," if (a) if we shift our focus away from individual members' actions toward the working of social and political institutions, and (b) adjust the relevant attitudes that are necessary for community in large-scale political communities.

On (a): rather than assessing personal interactions, the macro-level requires us to scrutinize values embedded in societal practices and institutions in terms of whether or not they express reciprocity, special concern, equal status etc. The relevant inequalities of standing are not, in the first instance, those between individual persons but between groups or categories of people based on salient social features. It is a minimum requirement of political community that these institutions' decision-making procedures give members equal opportunity for influence over what these rules are (Kolodny, 2014; Viehoff, 2014). Moreover, political community is impossible when the outcomes that such institutions produce deviate drastically from the ideals of communal sharing, reciprocity and mutual care (Schemmel, 2021).

On (b): imagine a society with just institutions, but citizens only obey laws because they are coercively enforced. We would not say that there exists community. This is because, as we saw, community requires members to have certain beliefs and attitudes. But how should we interpret these in the large-scale case? First, people will affirm formal institutions that express mutual concern, care, reciprocity, and so on (*public good orientation*) and they will not take advantage of the institutions' shortcomings (*civility*). But second, they will also affirm informal social norms and practices that express these attitudes. As Rawls says: "fraternity is held to represent a certain equality of social esteem manifest in various public conventions and in the absence of manners of deference and servility" (Rawls, 1999, p. 91). One plausible interpretation of what this element of community amounts to is a social or egalitarian ethos that governs people's social and economic interactions in their capacity as private citizens (Wolff, 1998; G. A. Cohen, 2008).

With these in mind, I want to suggest that whether a large-scale political society meets the ideal of community depends on three specific criteria:

1) the democratic quality of decision-making procedures for coercively-enforced public institutions (*democratic practices*);
2) the justice/fairness of outcomes produced by these public institutions (*institutional outcomes*); and
3) the quality of informal practices and norms and the extent to which they express a sense of unity, shared purpose with other members, equal standing, etc. (*social ethos*).

Now that I have stated the central elements of community for large-scale political societies, let me address one important question, namely its relation to justice. Both (2) and arguably (1) will form part of most philosophical accounts of justice. So on the view that I have sketched, issues of justice and community partly overlap. But that does not imply that justice and community are one and the same. This is the case for three reasons: First, valuable community depends on informal social norms and prevailing affective attitudes and emotions. Because we have no moral rights that others have such attitudes, they are not a matters of justice. Second, community may be less demanding than justice in that it can exist in scenarios where less than "full justice" is realized through public institutions: some measure of legitimacy in procedures and outcomes combined with the relevant attitudes may be sufficient for it. Third, the value of community is *collective*: it derives from integrated attitudes amongst members. As far as justice is concerned it matters *individually* for each person that her rights are respected, quite independently of other people's attitudes.

AI's Impact: Positive Elements

My aim is to analyze how AI is likely to affect political community. This section illustrates, in an optimistic spirit, how AI can positively contribute to community. To do so I assume the existence of a polity that satisfies (1) through (3) above; that is, there is a reasonably democratic decision-making procedure in place and public institutions guarantee a reasonably just distribution of social cooperation (e.g., through meaningful welfare state policies or a basic income). Finally, citizens do already, to a sufficient degree, have a shared social ethos that expresses identification, special concern, care, and so on.

Before assessing the potential positive effects of AI, let me specify what is meant by AI and "AI transformations." AI denotes both a *phenomenon* (that of intelligence displayed by non-biological agents) and a *field of human activity* (that of designing and using AI agents). AI agents are computerized machines that display some of the abilities standardly associated with human intelligence (e.g., the ability to learn from data and past "experience," the ability to pick strategies appropriate for the pursuit of one's aims, etc.). More informally, AI also refers to instances where specific *techniques* conducive to creating AI agents are at play (e.g., machine learning, automated data collection and analysis ("big data"), and robotics). Adopting this wider use, we can analyze the consequences of the growth in the use of machine-based decision-making and influence in various social domains, including

commerce, finance, healthcare, transportation, the provision of welfare and housing, education, law enforcement, and many others.

When we think about individual practices, AI may positively contribute either by securing elements of community already realized, or it could serve those aiming to change society in that direction for those aspects that still fall short. AI technologies (big data analysis, machine learning, etc.) have the capacity to better analyze and predict and, ultimately, to better control both the natural and the social world around us. How well we realize the ideal of community does, at least in part, depend on how well we succeed in understanding and controlling certain aspects of that social world. So AI technology may in principle be helpful for each of the individual constitutive components of the political community. It could improve political equality and the character of democratic decision-making; or it could enhance the core substantive outcomes of social, political, and economic institutions; or it could boost the informal aspects of community (e.g., public good orientation, civility, and social equality). Below I will offer examples how AI might make positive contributions to practices in each of these domains.

Democratic Practices

Let us start with the benefits that AI technology may hold in store for democratic decision-making. Collective democratic practices consist both of formal, institutionalized decision-making procedures and informal practices and norms (e.g., norms about truth-tracking and a willingness to change one's mind) (J. Cohen & Fung, 2021, p. 32; Himmelreich, 2021, pp. 5–7). AI technology can play (and has played) an important role in the accessibility of diverse standpoints and participants' ability to access information. Whilst the existence of information that one could consult is not tied to AI technology, their availability is, for it is only as a result of the sophisticated search algorithms on which today's online search engines run that citizens can access them. The same is true for participants' ability to disseminate views in the digital sphere. Without some algorithmic analysis/matching procedure, it is hard to see how each person's improved ability to publicize their views could translate into their being noticed by others. Equal access to and ability to disseminate information shapes both the formal and informal aspects of democracy: in respect of the first, it has the potential to address, to some extent, unequal opportunities to shape political outcomes. When some, but not all, can disseminate their views widely, this creates unequal *de facto* decision-making power. Likewise, when some, but not all, can easily verify whether some claim is true (e.g., whether a policy will advance their interests), then that too amounts to a de facto inequality of political power (Christiano, 2021, pp.8–9). AI-powered information technology may revert inequalities of this kind to some extent.

Moreover, complex matching algorithms and data analytics, once available at low costs to the democratic public, could improve citizens' ability to find and cooperate with others concerned about some particular issue and could, thereby, facilitate mass mobilization and create "counterpower" against powerful vested interests (Benjamin, 2019). Similarly, AI may facilitate the inclusion of voices that remain otherwise unheard for largely technical/efficiency reasons. For example, natural language processing/translation might help linguistic minorities to participate in public debates and it might simplfy some of the thorny issues relating to public representation and linguistic esteem in multi-language polities. With

regards to the informal norms governing democracy, AI technology, by drawing on new forms of data and their analysis, may make it easier to track the *common good*. Conversely, AI analysis may make it easier to systematically assess whether some agent has in the past been guided by some values, and whether they have respected norms of civility (Lorenz-Spreen et al., 2020).

Institutional Outputs

Next, consider AI's promise when it comes to the fair distribution of the benefits and burdens of social cooperation; that is, the part of community that overlaps with substantive justice or fairness. Whilst much has recently been written about the problems that automated, algorithmic predictions and decisions regarding the distribution of public benefits—and the metering out of public bads—have created (Eubanks, 2019), it is worth remembering that at least part of the initial impetus toward data-driven decisions was a dissatisfaction with the myopia, unreliability, and outright prejudice of human agents charged with distributing benefits and burdens (Kahneman et al., 2021). In principle, AI could, and frequently *does*, produce fairer and more efficient allocations of resources than those that average human counterparts produce, whether it concerns the domain of employment, housing, investment or the insurance (A. P. Miller, 2018). In all these domains AI could avoid waste, detect and correct for cognitive flaws inherent in human decision making, and also facilitate more equitable distributions of goods (Sunstein, 2019). Moreover, where there are trade-offs between values, including different interpretations of fairness, algorithms and automated data analysis may tell us how serious these are and how they may be lessened (Kleinberg et al., 2018).

Finally, sophisticated data analysis based on speech recognition and machine learning may become increasingly able to *uncover* patterns of bias, discrimination and inequity in the distribution of social benefits and burdens that public institutions generate (e.g., in law enforcement, public housing, and healthcare) (Voigt et al., 2017; Obermeyer et al,. 2019). Although we might already be aware of these inequities, large, quantified evidence of such phenomena may well be essential for gathering sufficient support to eradicate them.

Social Ethos

Perhaps one could think that on the third aspect of political community, namely the existence of informal social norms and a community-oriented social ethos that individuals share, it is hardest to think of ways in which AI could protect or improve such informal practices. But that is not so: even where such an ethos exists, political community faces threats to prevailing social norms that derive from unjust and discriminatory individual-level behavior. When such behavior becomes widespread, it will eventually change clustered beliefs around appropriate behavior. Here AI may be helpful.

As an example, consider norms about equal social status and how they can be undermined. Two frequently discussed phenomena here are implicit bias and micro-inequities/microaggressions. Implicit bias refers to prejudices and stereotypes that individuals hold without intending to do so or fully being aware of them (Brownstein, 2019). Micro-inequities

are small, unjust inequalities that manifest in interpersonal behavior. Such inequities, when they target disadvantaged groups, become micro*aggressions*. These, according to McTernan, "form a social practice that contributes to structures of oppression and marginalization" (2018, p. 269). The hope, recently put to the test by AI developers, is that AI will help us to reveal these previously hidden or unnoticed (micro)patterns of inequitable behavior, and it can provide strategies for preventing inequities from turning into equality-undermining and discriminatory social norms. A good real-world example here is online communication: Various researchers have recently presented microaggression detection algorithms that picks up subtle forms of put-down in twitter posts (Breitfeller et al., 2019; Ali et al., 2020). More ambitiously, we could think of AI-powered mobile devices that track one kind of barrier to equal social standing—implicit bias in everyday behavior that marginalizes socially salient groups—and reveal to members of society where their actions fall short of the relational ideal, thus enabling them to adjust their behavior.

AI-Transformed Practices: The Reality

The gist of the previous section was that because AI realizes an increase of our ability to shape the social world, it has, when used in the right way by the right actors, potentially positive consequences for realizing the ideal of community. This section suggests that, against the backdrop of societies that fall short of the ideal, the use of many AI technologies in specific practices has exacerbated rather than improved our ability to realize community. Before switching to a more structural long-term analysis of AI's impact in section five, here I demonstrate how AI has negatively impacted particular practices, again distinguishing between *democratic practices*, *institutional outputs*, and *the social ethos*.

To at least some extent all polities, even democratic ones, fall short of the ideal of community: Procedurally, democratic institutions are often marred by highly unequal de facto opportunities to influence public decisions. Substantively, benefits from social cooperation in many cases fail to satisfy any reasonable understanding of fair reciprocity. And to make matters worse, these institutional inequalities are almost always structured along socially salient group predicates like class, gender/sex, race, or sexual orientation. They thereby (re)produce discriminatory informal norms, (consciously or unconsciously held) biases, and entrench attitudes about unequal status and worth. Through oppression, marginalization and domination, informal norms harm individuals deemed inferior (Anderson, 1999, p. 312). But they also preempt the public good of community: social pathologies entrench social hierarchy, make it impossible to see the state as a joint project to realize justice, and render special concern and mutual care inappropriate amongst those harmed and those benefitting.

Democratic Procedures

The promise of AI technology, I said, lies in its potential to facilitate the curation and dissemination of information and to thereby improve deliberation and, through easier mobilization, equalize democratic power. The problem with AI, at least in the way it has so

far shaped democratic practices, lies in the fact that, on the one hand, it has not improved deliberation and, on the other hand, it has increased political inequality in some domains.

On the first score, it has been suggested that when it comes to information, "overabundance overloads policymakers and citizens, making it difficult to detect the signal amid the noise" (Dryzek et al., 2019). This interacts with and reinforces problems that flow directly from the fact that large swaths of citizens today gain information relevant for democratic choice through social networks. Given the overabundance of information, such networks necessarily apply sorting algorithms to what information each person is shown. As profit-maximizing enterprises, network providers will typically look for "engagement", such as news items that keep the user engaged with the site/app for the longest time. Two consequences that matter greatly for democratic deliberation have been noted: first, "newsworthiness" becomes tied more to viewers' emotional reactions in terms of approval or outrage than to actual importance and truth. AI technology thus contributes to the spreading of "fake news" and "post-truth" politics more generally. Second, sorting algorithms create "filter bubbles" such that, over time, individuals are presented more and more with information and standpoints that align with and further confirm their already-held beliefs. This makes good public deliberation harder (Watson, 2015; Nguyen, 2020).

Second, digitalization and AI technology have in fact dramatically increased the power to shape outcomes by some through the ability to influence opinions and beliefs of others by way of precisely determining what information they have. This should be obvious in the case of undemocratic regimes that control the online infrastructure and police newsmedia, social networks, and messaging services through the use of big data analytics and automated filtering (Diamond, 2019). But the use of big data and algorithms also shapes power in liberal democracies. First, it clearly offers those who control social media platforms unprecedented influence, whether they make use of it for their political purposes or not. Perhaps more importantly, big data analysis allows for "micro-targeting"—very fine-grained use of user data for political advertisement—and "hypernudging"—communication that nudges by constant updating with regards to what the voter responds to (O'Neil, 2016, p. 158). At least the latter technology exacerbates inequality not only by increasing power for those who can afford these technologies, but also by decreasing the power of those whose are, unbeknownst to them, manipulated by AI-powered nudges (Christiano, 2021, p. 4).

Distributive Outcomes

Recall that on the picture of community I have sketched, community requires a reasonably fair distribution of the benefits and burdens from social cooperation. When it comes to these, AI technology too often has not only failed to eradicate inequities, but has entrenched them. According to Eubanks, in the United States, the effect of the rollout of automation and algorithmic decision-making in the public domain has, on the whole, been to "divert the poor from public benefits, contain their mobility, enforce work, split up families, lead to a loss of political rights, use the poor as experimental subjects, criminalize survival, construct suspect moral classifications, create ethical distance for the middle class, and reproduce racist and classist hierarchies of human value and worth" (Eubanks, 2019, p. 183). This has been documented in various domains including criminal justice, healthcare, and social benefit provision (Eubanks, 2019; Mesquita, 2019). The fundamental underlying defects are

easy enough to grasp: AI agents or algorithms used to automate public policy are primed to make correct predictions (reach decisions) using mostly historical training data. Where either the data itself reflects unequal access to goods, or where the process of data collection is biased against the marginalized, or where either prediction or outcome variables need to rely on proxies that correlate with some socially salient group feature, the resulting decision will reproduce unjust outcomes (Mayson, 2018; Johnson, 2021).

Two further phenomena to highlight are *entrenchment* and *dispersal*. By *entrenchment* I mean to account for the fact that, as a result of automation, it becomes harder to criticize and contest inequitable institutional outcomes. There are various reasons for this. First, automated decision-making is often intransparent: the data sets used as well as the algorithms applied are frequently unavailable to those affected; many models and algorithms are proprietary; and even when they are not, the technology cannot easily be questioned by those affected, nor explained by those who execute it (Valentine, 2019, p. 367). Second, because they rely on mathematical models and machines, automated decisions seem like "objective" outcomes based on "rational" and "scientific" methods rather than human prejudice (Green, 2019, p. 70; Mesquita, 2019; Young et al., 2021). By *dispersal*, I mean the phenomenon, massively facilitated by big data and machine learning, that inequitable outcomes in one domain feed into public decisions in other domains, thus spreading injustice across public institutions. Consider one example described by Eubanks (2019, p. 144): as a result of unjust data collection processes, poor minority parents in urban areas are disproportionally likely to be wrongly targeted by algorithmic child-protection software. One variable that contributes to scoring badly is whether one has previously sought public assistance. Eubanks describes how parents have therefore become reluctant to access such facilities, knowing that it may increase the risk of having their child placed into foster-care.

Social Ethos

So far I have focused on AI's effect on inequitable outcomes of public institutions. But because these inequities are not randomly distributed across the population but heavily clustered among minorities and marginalized groups, they also matter greatly for whether or not people will develop the social and egalitarian ethos necessary for political community: AI-induced entrenchment and dispersal do not merely reproduce unjust outcomes, but they also enshrine forms of prejudice, stigmatization, and marginalization that condition private behavior and attitudes. And they make it inappropriate for those suffering from them to see the state as a joint project.

Prejudice in AI-decision-making is known to be prevalent far beyond decisions made by public authorities: similar issues have been documented in the application of AI in "private" commercial contexts, ranging from discrimination in algorithmic recruitment and assessment software—replication of human bias in training data (Barocas & Selbst, 2016)—over unjust inequalities in accuracy of face and body recognition software—mostly an instance of unjust data collection (Howard & Kennedy, 2020). These matter, inter alia, for security verification services and pedestrian safety in autonomous driving vehicles. In these domains, AI has been shown to entrench rather than limit social pathologies, such as community-inhibiting social practices. And as a result of data portability and network effects, these inegalitarian effects disperse throughout civil society and reinforce each other.

But beyond those instances where the application of AI technology entrenches or disperses already-existing social pathologies and thereby undermines communal attitudes, it can also *create* community-threatening cleavages. I will call this the phenomenon of *decommunitarization*. One way in which this can arise relates to a phenomenon already discussed in relation to the epistemic quality of democratic deliberation, namely the formation of filter bubbles and echo chambers as a result of algorithms: persistent sortition can lead to stratification of society into sub-level groups and, moreover, the formation of exclusive identities tied to these segmented groups. It is not only the implementation of AI in public institutions and practices relevant for democratic decision-making that may contribute to this phenomenon. Just as importantly, AI filtering may create hermetically closed circles of incompatible lifestyles, attitudes, and beliefs. Whilst not all of these will necessarily lead to decommunitarization (the liberal kind of political community described is compatible with very diverse conceptions of the good), disconnected lifestyles can lead to polarization; that is, a situation where segmented groups define themselves antagonistically (Chan et al., 2006; Baldassarri & Gelman, 2008).

Another way in which AI may have a decommunitarizing effect, this time on attitudes about mutual support and reciprocity, is what we might call "revealed competition" (Hussain, 2020). As an example, consider universal public health insurance where each citizen makes a roughly equal contribution (in terms of percentage of income, say) to receive access to healthcare. Such policies are arguably amongst the most straightforward expressions of community or solidarity. They communicate that, as a group, members determine access to essential health-related goods based on a principle of need, rather than wealth, prior conditions, irresponsible individual choice, and so on. Insurance of this kind is in-part popular because, from an *ex ante* perspective, most contributors are roughly symmetrically positioned. Perhaps we know about some pre-conditions, but, on the whole, we cannot differentiate between risk profiles with high precision. AI has the potential to fundamentally change this fact: where detailed knowledge about risk profiles are in fact available as a result of better data analysis tools, the likelihood that "wide reciprocity" practices break down increases; for those with better profiles now know that they could benefit from a pool that excludes others. So what AI effectively does is to "decommunitarize" practices by revealing possible conflicts of interests that, due to a natural veil of ignorance, had previously guaranteed mutual fate-sharing. And unfortunately, the phenomenon generalizes to all practices where previously unavailable information now reveals how prudential interests come into conflict.

Structural Changes: Datafication, Automation and a Disappearing Public Sphere

Looking at a range of practices, the previous two sections provided a mixed picture. On the one hand, AI technology has the potential to make positive contributions to community-relevant practices: it could strengthen democratic deliberation, improve the output of public institutions, and protect a community-oriented ethos. Yet on the other hand, its

implementation in imperfect societies has often reinforced existing patterns of inequity and exclusion that undermine this ideal. Moreover, we have seen that AI can also create new social cleavages.

I want to take a more structural perspective on what AI technology does to social and economic conditions. It seems clear that the development of productive forces in society (e.g., technology) stand in some important causal relationship to many social relations and practices. My conclusion for analyzing these is that although we can think of some "tech utopia" in which structural transformations make positive contributions to political community, the actual processes we see and their likely trajectory point in a less hopeful direction. Specifically, my focus is on three shifts: first, "datafication," and, relatedly, data capitalism. This AI-facilitated development changes central aspects of human sociality that directly bear on community. Second, "heavy automation," which is altering basic parameters of social cooperation and will redefine the meaning of work. Third, structural changes to the public sphere, which will alter core aspects of mass communication and will fundamentally reconstitute important pre-requisites of community.

Datafication/Data capitalism

Perhaps the biggest problem with the "optimistic picture" is that it makes it sound as if AI technologies were merely a tool or technique that can enhance our ability to pursue aims in certain domains. But AI also reconfigures ideational and material structures of social cooperation in a more profound way. Datafication is the process by which data of any kind is turned into a commodity. In the data economy; that is, the industry whose most important factors of production are data analysis technologies and the technical hardware to run them, surplus extraction occurs through the collection, analysis and sale (or curation and subsequent sale) of data. As far as user data is concerned, its economic value derives from its ability to allow producers and service providers to predict and shape behavior. The data economy is rapidly restructuring and reconfiguring relations of production in other domains of the economy, from manufacturing over finance to services. The rapid rise of this segment is seen as a structural transformation of capitalism. Talk of "data capitalism," "surveillance capitalism," or "informational capitalism" is meant to capture this radical change (J. E. Cohen, 2019; Zuboff, 2019).

Massive datafication of information about human persons shifts the relation in which individuals stand to each other in ways that we are only now beginning to grasp. First, the data economy creates a new class division between data controllers (and those able to pay for their services) and everybody else, and it grants those on top significant social control (Benn & Lazar, 2021, p. 10). A second crucial upshot is a drastic "horizontal" increase in interdependence when it comes to privacy and revealed information (Viljoen, 2021, p. 603ff.). The existence of large data warehouses and powerful data analysis tools means that data-collecting agent A's having some information about data subject S_1 dramatically increases what A can reasonably conclude about the predicates of some other data subjects S_2, S_3, and so on. Under datafication, there is no simple way to disentangle one agent's interest in revealing information and another agent's interest in not having information about them become public as a result of the former's actions (Viljoen, 2021, p. 605). But instead of democratizing collective choice about data use, processes of

"natural" monopolization—network goods are anti-rivalrous—lead to a situation where social power over data and its governance is increasingly accumulated in the hands of a few incredibly powerful corporations. Thus, "community guidelines" become more important than democratically enacted law. Datafication and its use in data capitalism, Viljoen suggests, is an "an unjust social process [that] denies individuals (both data subjects and those with whom they are in horizontal data relations) a say in the social processes of their mutual formation" (2021, pp. 641–642). Because community depends on social and political equality, the consequences of this structural transformation are disastrous.

Heavy automation

Although perhaps the most central one, the rise of data is only one amongst a number of structural shifts that AI introduces to economic cooperation and communal reciprocity practices. A second one that will likely play a fundamental role in the future stems from AI-facilitated automation of production, in which human labor, especially of the low- and medium-skilled variety, becomes increasingly dispensable. There are two aspects that are likely to exert a significant influence on the possibility of community in the future. First, and already visible in some domains, the automation of production is likely to further shift the balance between the respective share of the social product that go to labor and to capital[2]: If the capital share increases, the rich will move further apart from the professional middle class. Even though *some* material inequality is not necessarily community-undermining on a liberal conception, there is a legitimate concern that, where the lifestyle and habits of a small minority drastically differs from that of everybody else, community becomes hard to sustain.[3] Moreover, those lacking important resources will, with good reason, see their hardship as incompatible with attitudes of community (special concern, care, reciprocity) in light of the fact that the well-off could easily alleviate their suffering but fail to do so. A second upshot of heavy automation, although one that is disputed amongst economists, is that it will lead to massive redundancies and the inability of large swaths of members of society to participate in the labor market. Of course, in an optimistic spirit, we could imagine an AI-transformed, fully AI-automated economy. Such a system of production would obviate the need for any individual to seek employment in the form of wage labor as it would (fully automatically) allocate a fair share from social cooperation to each member of society, for example in the form of a universal basic income. One hope that some have formulated for such a "post-work society" is that it would eradicate many of the prevailing forms of social division and marginalization that derive from (involuntary) unemployment and poverty (Mason, 2016; Srnicek & Williams, 2016). The problem with this vision is that, given our current trajectory, it seems naively utopian. In reality, the disappearance of various types of employment would not lead to some unconditional provision, would thus have a dramatic material consequences on those lower-wage-earning segments of the labor market. This would lead to inferior opportunity sets for those whose labor is no longer needed. But, just as important from the standpoint of community, it would also, if current norms about the value of work in our lives persist, result in a loss of esteem and self-respect as structurally unemployed citizens will no longer be able to see themselves as productive, fully-contributing members of society.

Transformation of the Public Sphere

The final structural shift promoted by AI that I want to discuss does not, in the first instance, concern economic and material transformations but changes to the ideational structure of civil society. Although I have mentioned some AI-induced changes to democratic deliberation and social attitudes in earlier sections, I now want to argue that the interaction of a number of processes and phenomena amounts to more than alterations to specific practices: instead, AI-fueled digitalization is leading to a dramatic transformation of the public sphere, perhaps even its disapperance as a shared space of citizen deliberation. A functioning public sphere is one of political community's prerequisites. Thus, to the extent that AI-technology, together with other processes, is causing its disappearance, AI undermines community.

The public sphere is functionally defined: it is a domain in which democratic subjects encounter one another in a structured way outside of either the sphere of economic production or state authority in order to deliberate about the common good. A public sphere is well-functioning to the extent that citizens can and will engage in public deliberation and jointly construct social ideals. This requires an attitude of mutual *trust* amongst members, understood as a widely held, robust motivation to rely on others that is infused with normative expectations. Where such trust is lacking, it is unlikely that participants will adopt the necessary non-strategic attitude required for public deliberation. Moreover, a public sphere requires shared assumptions about political purposes and facts, something that, at least in large modern societies, seems difficult without mass media that present diverse views and structure and amplify political concerns (Habermas, 1991).

The digitalization of media together with AI technology causes the disappearance of functioning public sphere by fragmenting, polarizing, and commercializing the avenues for public debate and deliberation. First, it has resulted in the virtual extinction of one important source of civic deliberation and communication, namely local/regional engagement through subnational newspaper, TV stations, and so on. Moreover, as discussed, the online consumption of news, especially via social networks, has led to filter bubbles and echo chambers that make it exceedingly hard to rely on any shared assumptions in relation to either facts or values. This has massively contributed to fragmentation and polarization, which in turn eradicate social trust.

Might we not, in an optimistic mood, hope for some version of an "automated public sphere" or "data democracy," a technology-enhanced form of radical, direct democracy where public deliberation based on shared values and factual assumptions is unnecessary and policies are made algorithmically? As Susskind writes, "by gathering together and synthesizing large amounts of the available data—giving equal consideration to everyone's interests, preferences, and values—we could create the sharpest and fullest possible portrait of the common good" (Susskind, 2018, p. 247). Because everybody's preferences and values count equally in the aggregation, it would eradicate problems that plagued existing democracies (e.g., inequality of power, information, and status between citizens and their representatives, the exaggerated influence of the wealthy and the informed, etc.). And important for this section, it could obviate the very need for a functioning public sphere: all we need is our algorithms to tell us what is in our best interest. But this very idea of full automation contradicts the fundamental values of deliberative democracy and, derivatively, the

ideal community. For example, it would be a very impoverished notion of collective group agency: citizens would no longer together shape their common political life, but would only be recipients in a large distribution mechanism.

Conclusion

To conclude, I want to reflect on potential policies or strategies for addressing AI's dangers. First though, let me summarize. We started with a brief summary of community, and explained how its realization in large polities depends on working democratic practices, equitable distribution of the benefits and costs from social cooperation, and, importantly, informal aspects of how people relate to each other (social ethos). We then investigated how AI shapes individual community-relevant practices and institutions.

Although we saw a hopeful account of what AI technology could do for particular community-related practices in section three, it was also shown how AI's implementation has had significantly detrimental effects on community in existing societies that contain social pathologies such as marginalization and discrimination. I also explained how AI may give rise to *new exclusionary social categories* through decommunitarization. I suggested that we would be missing something fundamental if we only looked at the use of AI technology in specific practices. Instead, the rise of data capitalism, and the structural changes it creates for other domains of the social system of production and deliberation are crucial for how the ideal of community can be realized. We concluded that there is much less to celebrate here: through datafication, AI contributes to inegalitarian social relations, and through heavy automation, it has the potential to undermine practices of productive reciprocity in the economy. Finally, the combination of digitalization of mass media and AI technology has led to a structural transformation of the public sphere such that some essential elements, such as social trust, cohesion, and a common basis for public deliberation, are increasingly absent.

So should we give up any hope of realizing community in the age of AI? I want to suggest that we should not. Instead, we should look to a mix of strategies and policies tailored to our complex predicament. Some of the problems, I want to suggest, can be addressed with well-known interventions. Others will require new ones that normative theorists should focus their efforts on developing.

Let me begin with some of the more straightforward points: with regards to some of the examples I have described, AI's implementation threatens community because the way in which AI is used creates injustices, and injustice forecloses (valuable) community. Whilst it may be technologically challenging to operationalize a system that harnesses big data and modern information technology without perpetuating past injustices through training data that contains it, there is relatively little that normative philosophical analysis can contribute other than to suggest that it is wrong to implement AI-run systems that perpetuate injustice, either in public services or in private economic contexts. But the right solution to unjust AI-system not always, or even typically, not to make any use of AI. Rather, the answer will often have to be more equitable, more transparent and more explainable algorithms.

Second, some of the AI-induced macro-level effects that negatively shape our ability to realize community, whilst clearly challenging in terms of their scale, do not seem unprecedented or as necessitating fundamentally new strategies. Take, for example, the two consequences attributed to AI-induced heavy automation in the economy, namely redundancy and top-level inequality: As far as redundancy is concerned, it seems that received public policy strategies like active state involvement in the process of skill formation and (re-) education, combined with robust social insurance mechanisms could address this—just like they have, to some extent, been able to preserve community through earlier waves of technological-cum-economic transformation. And as far as AI-facilitated run-away inequality at the top-level is concerned, the best policy instruments by far seems good old-fashioned taxation, although one that is globally coordinated and focused on AI-derived surpluses.

Designing appropriate policy responses is much harder when we think about datafication and AI-induced shifts in public deliberation. Even though it is clear that the jarring new inequalities of social power that have arisen through the data economy and network-operating corporations are incompatible with community, it is difficult to conceive of normatively appealing, feasible alternatives.

Perhaps one important insight on which we may build is this: even if we cannot and arguably should not reverse the increase in our ability to predict and control the natural and social world around us that AI creates, it does make a very significant difference whether these new forms of social power are exercised *individually* and for the purpose of private economic gain, or *collectively* through shared democratic agency and in the spirit of the common good. One complication our current predicament is that the thing that is needed—shared democratic agency to collectively control and manage AI's transformative potential—is itself endangered by AI-induced social stratification and the disappearance of shared spaces for citizen deliberation. Normative theorizing does not, at the moment, offer much by way of meaningful guidance for the kinds of institutions that would express core elements of community like social and political equality, reciprocity and a social ethos. It is here that innovative reflection from both normative theorists and analysts of the data economy is urgently needed.

Acknowledgements

For helpful comments on earlier drafts, I thank Johannes Himmelreich, Daniel Viehoff, and Iason Gabriel.

Notes

1. There is no generally agreed term for the phenomenon I have in mind. People addressed the topic using "fraternity" (Rawls, 1999, pp. 90–91), "community" (G. A. Cohen, 2009, pp. 34–37; 2008, p. 43; D. Miller, 1989, p. 234), or "solidarity" (Shelby, 2005, pp. 67–71; Nagel 1991, p. 178). I use *community* throughout.
2. See Carles Boix, "AI and Democracy" in this handbook.
3. This was the kind of division G.A. Cohen was at times worried about (2009, pp. 35–36).

References

Ali, Omar, Scheidt, Nancy, Gegov, Alexander, Haig, Ella, Adda, Mo, & Aziz, Benjamin. (2020). Automated detection of racial microaggressions using machine learning. *Proceedings of IEEE SSCI 2020*, September, 2477–2484. http://www.ieeessci2020.org/index.html.

Anderson, Elizabeth. (1999). What is the point of equality? *Ethics 109*(2), 287–337.

Baldassarri, Delia, & Gelman, Andrew. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology 114*(2), 408–446. https://doi.org/10.1086/590649.

Barocas, Solon, & Selbst, Andrew D. (2016). Big data's disparate impact essay. *California Law Review 104*(3), 671–732.

Benjamin, Ruha. (2019). *Race after technology: Abolitionist tools for the new Jim code*. 1st ed. Polity.

Benn, Claire, & Lazar, Seth. (2021). What's wrong with automated influence. *Canadian Journal of Philosophy* First View, 1–24. https://doi.org/10.1017/can.2021.23.

Breitfeller, Luke, Ahn, Emily, Jurgens, David, & Tsvetkov, Yulia. (2019). Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. EMNLP-IJCNLP 2019, 1664–1674. https://doi.org/10.18653/v1/D19-1176.

Brownstein, Michael. (2019). Implicit bias. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Chan, Joseph, To, Ho-Pong, & Chan, Elaine. (2006). Reconsidering social cohesion: Developing a definition and analytical framework for empirical research. *Social Indicators Research 75*(2), 273–302. https://doi.org/10.1007/s11205-005-2118-1.

Christiano, Thomas. (2021). Algorithms, manipulation, and democracy. *Canadian Journal of Philosophy*, November, 1–16. https://doi.org/10.1017/can.2021.29.

Cohen, G. A. (2008). *Rescuing justice and equality*. Harvard University Press.

Cohen, G. A. (2009). *Why not socialism?* Princeton University Press.

Cohen, Joshua, & Fung, Archon. (2021). Democracy and the digital public sphere. In Lucy Bernholz, Landemore, Helene, and Bob Reich (Eds.), *Digital technology and democratic theory* (pp. 23–61). University of Chicago Press.

Cohen, Julie E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press Inc.

Diamond, Larry. (2019). The road to digital unfreedom: The threat of postmodern totalitarianism. *Journal of Democracy 30*(1), 20–24. https://doi.org/10.1353/jod.2019.0001.

Dryzek, J.S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J.N., Felicetti, A., Fishkin, J.S., Farrell, D.M., Fung, A., Gutmann, A., & Landemore, H. (2019). The crisis of democracy and the science of deliberation. *Science 363*(6432), 1144–1146. https://doi.org/10.1126/science.aaw2694.

Dworkin, Ronald. (2000). *Sovereign virtue: The theory and practice of equality*. Harvard University Press.

Eubanks, Virginia. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Reprint Edition. Picador.

Fiske, Alan P. (1993). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review 99*(4), 689–723. https://doi.org/10.1037/0033-295X.99.4.689.

Green, Ben. (2019). *The smart enough city: Putting technology in its place to reclaim our urban future*. The MIT Press.
Habermas, Jurgen. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press.
Himmelreich, Johannes. (2021). Should We Automate Democracy? In Carissa Véliz (Ed.), *In The Oxford Handbook of Digital Ethics*. Oxford University Press.
Howard, Ayanna, & Kennedy, Monroe. (2020). Robots are not immune to bias and injustice. *Science Robotics 5*(48), 1–2. https://doi.org/10.1126/scirobotics.abf1364.
Hussain, Waheed. (2020). Pitting people against each other. *Philosophy & Public Affairs 48*(1), 79–113.
Johnson, Gabbrielle M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese 198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y.
Kahneman, Daniel, Sibony, Olivier, & Sunstein, Cass R. (2021). *Noise: A flaw in human judgment*. William Collins.
Kleinberg, Jon, Ludwig, Jens, Mullainathan, Sendhil, & Sunstein, Cass R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis 10*, 113–174. https://doi.org/10.1093/jla/laz001.
Kolodny, Niko. (2014). Rule over none II: Social equality and the justification of democracy. *Philosophy & Public Affairs 42*(4), 287–336.
Lorenz-Spreen, Philipp, Lewandowsky, Stephan, Sunstein, Cass R., & Hertwig, Ralph. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour 4*(11), 1102–1109. https://doi.org/10.1038/s41562-020-0889-7.
Margalit, Avishai, & Raz, Joseph. (1990). National self-determination. *Journal of Philosophy 87*(4), 439–461.
Mason, Paul. (2016). *PostCapitalism: A guide to our future*. 1st ed. Penguin.
Mayson, Sandra G. (2018). Bias in, bias out. *Yale Law Journal 128*(8), 2218–2301.
McTernan, Emily. (2018). Microaggressions, equality, and social practices. *Journal of Political Philosophy 26*(3), 261–281. https://doi.org/10.1111/jopp.12150.
Mesquita, Ethan Bueno de. (2019). The perils of quantification. *Boston Review*. https://bostonreview.net/forum_response/ethan-bueno-de-mesquita-perils-quantification/.
Miller, Alex P. (2018). Want less-biased decisions? Use algorithms. *Harvard Business Review 26*(26). https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.
Miller, David. (1989). *Market, state, and community*. Clarendon Press.
Nagel, Thomas. (1991). *Equality and partiality*. Oxford University Press.
Nguyen, C. Thi. (2020). Echo chambers and epistemic bubbles. *Episteme 17*(2), 141–161. https://doi.org/10.1017/epi.2018.32.
Obermeyer, Ziad, Powers, Brian, Vogeli, Christine, & Mullainathan, Sendhil. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.
O'Neil, Cathy. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
Rawls, John. (1999). *A theory of justice*. Revised 1999. Belknap Press of Harvard University Press.
Schemmel, Christian. (2021). *Justice and egalitarian relations*. Oxford University Press.
Shelby, Tommie. (2005). *We who are dark: The philosophical foundations of black solidarity*. Harvard University Press.
Srnicek, Nick, & Williams, Alex. (2016). *Inventing the future: Postcapitalism and a world without work*. Verso Books.
Sunstein, Cass R. (2019). Algorithms, correcting biases. *Social Research 86*(2), 499–511.

Susskind, Jamie. (2018). *Future politics: Living together in a world transformed by tech*. Oxford University Press.

Tönnies, Ferdinand. (2002). *Community and society*. Translated by Charles Price Loomis. David & Charles.

Valentine, Sarah. (2019). Impoverished algorithms: Misguided governments, flawed technologies, and social control. *Fordham Urban Law Journal 46*(2), 364–427.

Viehoff, Daniel. (2014). Democratic equality and political authority. *Philosophy & Public Affairs 42*(4), 337–75. https://doi.org/10.1111/papa.12036.

Viljoen, Salome. (2021). A relational theory of data governance. *Yale Law Journal 131*(2), 573–654. https://doi.org/10.2139/ssrn.3727562.

Voigt, Rob, Camp, Nicholas P., Prabhakaran, Vinodkumar, Hamilton, William L., Hetey, Rebecca C., Griffiths, Camilla M., Jurgens, David, Jurafsky, Dan, & Eberhardt, Jennifer L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences 114*(25), 6521–6526. https://doi.org/10.1073/pnas.1702413114.

Watson, Jamie Carlin. (2015). Filter bubbles and the public use of reason: Applying epistemology to the newsfeed. In F. Scalambrino (Ed.), *Social epistemology and technology: Toward public self-awareness regarding technological mediation* (pp. 47–57). Rowman & Littlefield.

Weber, Max. (2002). *Wirtschaft und gesellschaft. Grundriss der verstehenden soziologie*. 5th edited edition. Mohr Siebeck.

Wolff, Jonathan. (1998). Fairness, respect, and the egalitarian ethos. *Philosophy and Public Affairs 27*(2), 97–122.

Young, Matthew M., Himmelreich, Johannes, Bullock, Justin B., & Kim, Kyoung-Cheol. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4*(3), 244–58. https://doi.org/10.1093/ppmgov/gvab006.

Zhao, Michael. (2019). Solidarity, fate-sharing, and community. *Philosopher's Imprint 19*(46), 1–13.

Zuboff, Shoshana. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power*. Profile Books.

SECTION III

DEVELOPING AN AI GOVERNANCE REGULATORY ECOSYSTEM

VALERIE M. HUDSON

CHAPTER 13

TRANSNATIONAL DIGITAL GOVERNANCE AND ITS IMPACT ON ARTIFICIAL INTELLIGENCE

MARK DEMPSEY, KEEGAN MCBRIDE, MEERI HAATAJA, AND JOANNA J. BRYSON

Introduction

The continual process of societal digitalization has led to a dynamic and almost amorphous regulatory environment. Compare the regulation of digital technologies to that of say, pharmaceutical drugs, consumer-grade food products, or investment instruments, and it becomes clear that the degree of regulatory understanding and policy precision differs widely, even though digital technologies can have just as much impact on human cognitive and neurological capacities, health, and development as our medicine, food, and finances. One of the most important digital-technology concepts currently at play within this dynamic is that of Artificial Intelligence (AI). Having a clearly defined regulatory environment for AI would be helpful for developing not only widely shared understandings of the technology, but more urgently to ultimately enable governments to protect their citizens' interests and coordinate at a global scale to address issues that cannot be contained by national borders.

One leading model of transnational coordination is the European Union (EU), a political and economic union of nations with harmonized law only specifically where member nations agree by treaty to so harmonize. Founded in 1993, the zone of the EU's partially shared currency (the euro), has become persistently one of the world's three largest economies. The region is also taking increasing leadership roles in global issues such as sustainability, health, and climate.

AI regulation has become a priority for the EU's executive branch of government, the European Commission ("the Commission"). In her recent first speech before the European Parliament, the new president of the Commission, Ursula von der Leyen, committed to

adopting "a coordinated European approach on the human and ethical implications of artificial intelligence" (Von der Leyen, State of the Union, 2020, p. 13).

In today's digitalized world, transnational dependencies are increasingly common, and therefore transnational regulatory and governance frameworks are needed that take these dependencies into account. The way that the EU is handling digital regulation generally, and AI in particular, is of great interest because these regulations chart new territory in the digital technology space that likely provides valuable learnings to governments globally. Since the EU is a trading bloc of independently functioning and historically warring nations who now "harmonize" their legislation to create unified market policies, its mechanisms for building consensus and coordination across a multiplicity of cultural and governance structures is particularly likely to be both useful and even instructive. Further, in addition to the trans-national, but within-jurisdiction nature of EU legislation, academics have recently increasingly examined extra-territorial impacts of EU law, as what has become known as the "Brussels Effect" (Bradford, 2020). This and similar effects for other regions are an important further focus of this chapter.

With all of this in mind, this chapter charts the most salient attempts to date in providing global and transnational governance frameworks for AI worldwide, but with an emphasis on existing EU regulatory acts and proposals. Such acts include the General Data Protection Act which aims at digital privacy widely, including for AI. Proposals include the Digital Services Act (DSA) focusing on impacts and remedies for familiar platform AI, such as Websearch and social media, and the EU Commission's more recent Artificial Intelligence Act (AIA),[1] which focuses on identifying and regulating potentially hazardous applications of the technology. We start first with a brief definition of terms as they will be used in this chapter and follow with an overview of existing transnational AI governance efforts. The ensuing sections delve deeper into the EU regulatory process, its development over time, and its culmination in the Brussels Effect of EU law, which occurs upon the fulfilment of several conditions. The closing sections of the chapter assess two of the major EU regulatory proposals governing AI: the Digital Services Act (DSA) and the AI Act (AIA).

Defining AI, and Context

Definitions of AI abound, varying in their technical precision, legal utility, and prescriptive versus descriptive quality. AI is all around us (Bryson & Wyatt, 1997) and, like other systems, it can often be made out to be more complicated than need be. Unfortunately, such convolution in particular regulatory contexts may often be deliberate, aiming to advance an agenda or bypass potential scrutiny. In addition to this, AI remains a contested concept also given its universality as a general-purpose technology (Brundage & Bryson, 2016). Thus, to date, there has been difficulty when it comes to reaching consensus on a single definition of AI. This dispersed state of definitional affairs increases confusion and has negative implications for the regulation and governance of AI.

It is important to examine the interests and ambitions driving certain definitions of AI with a critical lens. For example, the Organization for Economic Co-operation and Development (OECD) has a *trade and economic progress mandate*; this is reflected in their definition, which it published with its OECD principles on AI.[2] According to the OECD,

an "Artificial Intelligence (AI) System is a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (OECD, 2019, p. 7).

Such a definition can be contrasted to that offered by the EU's proposed AIA which states that:

> "artificial intelligence system" (AI system) means software that is developed with one or more techniques and approaches listed in Annex 1 (of the AIA)[3] and can, for a given set of human-defined objectives, generate outputs, such as content, predictions, recommendations, or decisions influencing the environments they interact with.
>
> (European Commission, AIA, 2021, p. 2)[4]

These definitions are broad, but for the purpose of this chapter we will use an even broader and more succinct definition for AI: "AI is any artifact that extends our own capacities to perceive and act" (Bryson, 2019). Although it is an unusual definition, it might, as Bryson notes, "also give us a firmer grip on the sorts of changes AI brings to our society, by allowing us to examine a longer history of technological interventions" (Bryson, 2019). Further, it does so while capturing the essential core of the previously listed definitions—producing or altering action in a context via an artifact—while avoiding the hazard of trying to list (even in a relatively easy-to-update appendix) all the means by which this might be achieved.

An Overview of Existing Transnational AI Governance Efforts

Transnational AI governance is becoming increasingly important, and research on the topic has begun to rapidly proliferate, especially research focusing on the regulation and global coordination of such regulation for technological advances (Erdelyi et al., 2018; Deeks, 2020; Crootof et al., 2020; Beaumier et al., 2020). However, many scholars fail to sufficiently acknowledge or discuss pre-existing AI regulatory policy, which has evolved over several decades. Research and deployment of AI has, so far, been primarily "up-regulated" with very significant government and other capital investments (Miguel & Casado, 2016; Technology Council Committee on Technology, 2016; Brundage & Bryson, 2016; cf. Bryson 2019).

In this context, there are two important points to note:

- No one is talking about introducing regulation to AI. AI already exists and has always existed in a regulatory framework (Brundage & Bryson, 2016; O'Reilly, 2017). What is being discussed is whether that framework needs optimizing.
- Regulation has so far mostly been entirely constructive, with governments providing vast resources to companies and universities developing AI. Even where regulation constrains, informed and well-designed constraints can lead to more sustainable and even faster growth.

In this regard, it is possible to draw comparisons to the finance sector. Finance has always been regulated, but, as the global financial crisis (GFC) of 2007–2009 demonstrated,

regulations must be overhauled and optimized. We are at similar cross-roads with AI. As the technologist Benedict Evans argues: "[Information] Tech has gone from being just one of many industries, to being systemically important to society" (Evans, 2020). If something is "systemically important to society," it must be governed and regulated. Indeed, day-to-day life is becoming increasingly intertwined with AI. The welfare of society and citizens may be influenced by decisions that are becoming increasingly made by algorithms. Such changes have led to extensive research (e.g., AlgorithmWatch, 2020; EU Fundamental Rights Agency, 2020) and a further drive for "downwards regulation"—constraint—not least where privacy, surveillance, bias, and discrimination are concerned.

The lack of any formal regulatory structure to address AI concerns on a transnational level, and to legally hold corporations (including "platforms") and governments to account, has led to the rapid emergence of international governance and ethics fora, political fora, and standards developing efforts. Between 2016 and 2019, for example, more than 80 AI ethics documents—including codes, principles, frameworks, and policy strategies—have been produced by corporations, governments, and NGOs (Schiff et al., 2019). Perhaps the OECD Principles on AI come closest to a global consensus with 42 signatory states.[5] The core five OECD principles were soon subsequently adopted as the G20 AI Principles,[6] bringing the total number of signatory states to 50. The OECD Principles, combined with the United Nations Sustainable Development Goals, are now core frameworks for all efforts by the Global Partnership on AI, another recent transnational AI "initiative."

The Global Partnership on AI (GPAI), just mentioned, is another example of a recent high-profile initiative in AI governance. Launched in June 2020, it consisted at initiation of 15 partners.[7] In December 2020, four additional nations were admitted, and UNESCO became a member of the partnership as an observer. The GPAI is intended to be the preeminent global forum where working groups of chosen AI experts and practitioners meet to discuss and inform policy in a multidisciplinary approach. The present goal is for the GPAI to "support and guide the responsible adoption of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals" (GPAI, November 2020). Interestingly, the GPAI is not explicitly portrayed as a governance mechanism; in fact, its terms of reference prohibit "normative" outcomes. Nevertheless, in establishing a partnership developing shared governmental capacities in the cooperative deployment of AI interventions, it has clear impacts and consequences for AI transnational governance. Coinciding with the development of different transnational organizations, studies are also becoming increasingly available that map the content of existing principles and guidelines for AI regulation around the globe. Such studies aim to understand where a global agreement on AI guidelines is emerging or indeed has emerged (Jobin et al., 2019; Lorenz, 2020). Such research suggests growing global convergence around five ethical principles—transparency, justice and fairness, non-maleficence, responsibility, and privacy and defense of the human individual (Jobin et al., 2019; Bryson, 2017).

These principles correspond well to principles retained within the OECD and G20 approaches to AI, to which most wealthy nations are signatories, including all members of the EU, the United States, and China. Nevertheless, some argue that substantive divergences exist in how these principles should be interpreted and applied. Moreover, they question the usefulness of guidelines and principles. Principles tend to act a means of formalizing non-binding guidelines for national governments to abide by and they remain attractive to

"condense complex ethical considerations or requirements into formats accessible to a significant portion of society, including both the developers and users of AI technology" (Stix, 2021, p. 2) as well as by providing a useful starting point from which to develop more formal standards and regulation (Whittlestone et al., 2019). Principles are sometimes seen as too high level, difficult to practically implement, and particularly prone to manipulation—especially by industry (Rességuier et al., 2020). Nevertheless, the OECD has a good history of first expending effort gaining consensus around principles, and then seeing them implemented as law by member nations. This process of legislation is being followed now by several institutions, notably the EU.

EU Regulatory Process

That the EU has led the way with proposals to legislate AI and digital technology more widely should not necessarily be a surprise. The development of the Digital Single Market has long been recognized as a priority by the European Commission's Digital Single Market Strategy (DSM)[8]—such strategy has formed an important piece of President von der Leyen's Commission agenda for Europe 2019–2024.[9] Digital regulation in the EU, including efforts to regulate AI, are first and foremost about the need to ensure the efficient functioning of the Single Market. But this functioning must be based on the four freedoms: of goods, services, capital, and labor. Additionally, the "European Approach" to digital governance aims to ensure that development is done in a way that is consistent with the EU values of human dignity, freedom, democracy, equality, rule of law, and human rights.[10]

In the past, the EU has been accused of regulatory activism and, most notably by former President Obama, as being protectionist, but there is little evidence to support this claim. In fact, recent judgments when analyzed suggest the opposite (Bradford, 2020, p. 104). Rather, the EU is simply a tough regulator, whether against foreign or domestic firms. EU citizens and NGOs have demanded more protective regulations and have been vocal in demanding more stringent consumer, environmental, and health protections than their counterparts in other parts of the world. It is also well known that EU citizens are particularly distrustful of the conduct of dominant companies in the digital realm and are more concerned about the integrity of their personal data which in turn has led to regimes such as the GDPR (Pfeiffer et al., 2021).[11]

The EU's tendency to harmonize[12] standards upwards has also been facilitated by changes to EU Treaties that have enabled regulations and directives to be adopted with a qualified majority[13] of the Council, as opposed to unanimity. The move towards qualified majority voting (QMV) goes back to the adoption of the Single European Act (SEA),[14] which came into force in 1987 and paved the way to completing the Single Market by 1993, and the four freedoms as mentioned previously. The ability to proceed with legislation, even in the absence of consensus, established the foundation for significant rulemaking in the aftermath of the SEA. Had member states insisted on unanimity as the default decision-making rule, it is doubtful that the ambitious regulatory agendas that have been undertaken by the EU over the last 30 odd years would have emerged (Bradford, 2020).

Such ambitious regulatory agendas have driven the development of extensive scholarship, which aims to understand the true regulatory reach of European institutions, and the

emergence of the EU as a, or even the, principal shaper of global standards across a wide range of areas (Bradford, 2012; Schwartz, 2019; Scott, 2014; Bradford, 2020). Often called "soft power with a hard edge" (Goldthau & Sitter, 2015), the EU makes up for its lack of hard power through its use of policy and regulatory tools. Compliance with EU policy is motivated by access to its enormous market. Often, as in the case of the general data privacy regulation (GDPR), the same regulatory acts not only set out measures to protect individuals within that market, but also to smooth and harmonize access to the full market, with the ultimate goal of a net positive increase in both commerce and other aspects of well-being.

The "Brussels Effect," a term coined by the EU law scholar Anu Bradford, refers to the EU's unilateral ability to regulate the global marketplace. This can occur *de facto* when global corporations respond to EU regulations by adjusting their global conduct to EU rules. No regulatory response by foreign governments is needed; corporations have the business incentive to extend the EU regulations to govern their worldwide production or operations simply because of the EU's own economic strength when taken as a unitary actor, combined with the utility of simplicity in maintaining a relatively unified product offering world-wide.

The Brussels Effect is considered *de jure* when the governments of third countries (those countries outside the EU) adopt EU-style regulations after global corporations have adapted their global conduct rules to conform to EU rules. In this case, companies often have the incentive to lobby their governments for higher standards in their home jurisdictions. This ensures that they are not at a disadvantage when competing domestically against companies that do not export to the EU and that, therefore, have no incentive to conform their conduct or production to potentially costly EU regulations (Bradford, 2020).

Bradford (2020) asserts that for the Brussels Effect to occur, five conditions need to be cumulatively met. However, follow-up research has found that in practice it can occur in the absence of some criteria being fulfilled (Preusse et al., 2020). Bradford's five conditions for the Brussels Effect are sufficient market size, regulatory capacity, stringent standards, inelastic targets, and non-divisibility.

Market Size

The EU Single Market is the world's largest single market area in which goods can move freely. It is also a place where the Commission's goal of high safety standards and the protection of the environment is maintained by stringent regulations. It accounts for 450 million consumers and 22.5 million small- and medium-sized enterprises (SMEs).[15] As Bradford notes, "Large market size is, indeed, a pre-condition for unilateral regulatory globalization. Yet the jurisdiction must also possess sufficient regulatory capacity to exercise global regulatory authority."[16] The EU at least is largely able to regulate its own markets in terms of the products its citizens and residents are exposed to.

Regulatory Capacity

The EU has institutional structures in place that can adopt and enforce regulations effectively[17] and which themselves reflect the preferences of its key stakeholders, EU citizens. But what sets the EU apart is the depth of its regulatory expertise and resources with which

it can enforce its rules and thereby exert its authority over market participants—within or outside of its jurisdiction. In the case of non-compliance, the EU imposes sanctions. Only jurisdictions with the capacity to impose significant financial penalties on others (as the EU often has) or exclude non-complying firms from their market can force regulatory adjustment (Bradford, 2014).

Inelastic Targets

These targets do not refer to the traditional usage of inelasticity in economics (sensitivities of demand when variables such as price change), but rather to the products or producers that are regulated by the EU. In the case of the Brussels Effect, the location of the consumer within the EU, as opposed to the location of the manufacturer, determines the application of EU regulation to the targeted product. Union geography is largely inelastic but can alter with Union membership.

Stringent Regulations

Studies (Guasch & Hahn, 1999; Elliott et al., 1995, as cited in Bradford, 2020) show that wealthier states are more likely to have higher domestic demand for stringent regulations. Beyond the EU's use of a stringent regulatory agenda as a tool for promoting further integration, there are two further crucial factors that also differentiate the EU from other states and regions in its demand for tough regulations (Bradford, 2020):

- Europe's greater belief in government rather than markets as the means to generate fair and efficient outcomes (ideology) (Pew Research, 2018).
- An environment in Europe where higher relative importance is given to public regulation over private litigation and a lower threshold for intervention by regulators in case of uncertainty.

Note that this "greater" belief in government is relative to the apparent trade-offs made in some other large regulatory regions, but not necessarily all. No markets are entirely free of regulation, nor is any government fully controlling its economy. What varies is the extent of reliance on these two strategies. Also note that differences in beliefs concerning such trade-offs may be justified by real regional variation in governing capacity.

Non-divisibility

This refers to the practice of standardizing—in this case when a company standardizes its production or business practices across jurisdictions. Bradford (2020) differentiates between "legal non-divisibility," when legal requirements and remedies drive uniform standards (for example in the case of a merger where rules of the jurisdiction with toughest competition laws prevail), and "technical non-divisibility," which refers to the difficulty of separating the firm's production or services across multiple markets for technological reasons. Either form increases regulatory power, but the EU's special concerns and legislative efforts of course most affect the former.

The EU Commission's Presently Proposed Regulations: The Digital Services Act (DSA) and the AI Act (AIA)

The Digital Services Act (DSA)[18] and the Commission's proposal to regulate AI systems (AIA, for AI Act)[19] were published on December 15, 2020 and April 21, 2021, respectively. At the time of writing, both proposals have given rise to global conversations on how to regulate the large technology platforms (which are ostensibly the focus of the DSA) and the use of digital automation by public sector authorities (arguably one of the main targets of the AIA). The reaction to both these proposals is not unlike the wave of debate about the nature of data privacy that followed the enactment of the GDPR, and which continues today. This section provides a perspective on both proposals but starts by regarding privacy and its status within European digital legislation.

The significance of privacy as a fundamental right

It is important to understand the significance of privacy in Europe as a core component of the EU's digital legislation. The philosophy behind the EU's fundamental rights approach to privacy is to foster self-determination of individuals by granting them enhanced control over their personal data (Lynskey, 2020). This philosophy can be traced back to the European Convention on Human Rights (ECHR),[20] a treaty document which guarantees a fundamental right to privacy, drafted not by the EU but by the longer established and significantly larger Council of Europe. All EU member states are among the 47 signatories to the ECHR, making all Europe beneficiary of its privacy rules. However, it was not until 2009 that the Lisbon Treaty gave legal force to the Charter, despite it having been "solemnly proclaimed" by the European Parliament, Council of Ministers, and the European Commission (but not the Member States) at Nice in December 2000 (Anderson et al., 2011). At the time it lacked the binding force of law, but as Anderson notes:

> The Convention members had decided to proceed on the basis that the Charter should be drafted "'as if' it might be legally binding," and further "its 54 Articles—which are for the most part replicated in the Charter as finally adopted in 2007—were thus drafted with an eye to the possibility of judicial enforcement.
>
> (Anderson et al., 2011, p. 10)

Similarly, the GDPR also furthers the goal of protecting EU citizens, their data, and their privacy. Privacy scholar, Paul Schwartz (2019) put the impact of the GDPR as follows:

> Proof of the influence of the GDPR and EU data protection law goes beyond the hefty sums spent by U.S. companies to comply with them. The EU has taken an essential role in shaping how the world thinks about data privacy. Even corporate America draws on EU-Centric language in discussing data privacy.
>
> (Schwartz, 2019, p. 3)

The EU frames the extraterritorial reach of the GDPR and justifies the measures it has taken to enforce compliance with the law, because they are necessary to ensure the continued protection of fundamental rights of EU residents when their data are processed beyond EU borders. This reasoning is echoed in the DSA and the AIA.

The Digital Services Act (DSA)

The DSA and the AIA proposals represent a seismic shift in how the EU regards digital policy, moving from voluntary codes to legally binding regulatory structures to ensure more accountability, transparency, and responsible behavior from market participants. Over the last 20 years there has not been a meaningful attempt to "regulate the Internet".[21] This is now changing.

The DSA builds on the existing e-Commerce Directive,[22] which was adopted in 2000, and was the first attempt by the EU to set up a framework to remove obstacles to cross border online services in the Internal Market. The e-Commerce Directive is acknowledged as being the first piece of EU regulation that set clear limits on liability for digital platforms—meaning that they were not held responsible for illegal material uploaded to their sites, but rather only responsible for bringing down illegal material when informed. In other words, without this legal safe harbor, the Internet would likely not have grown to what it is today, as this would have restricted the publishing of user-generated content (Echikson, 2017).

The e-Commerce Directive has also had important consequences in allowing the digital culture in Europe to evolve. It enabled digital firms in one EU member state to do business in another without any restrictions and it stipulated, importantly, that when they do so, it is the "country-of-origin rules" that apply for most purposes. This remains the case with the DSA. The "country-of-origin" principle is a key principle that was laid out in the e-Commerce Directive whereby providers of online services are subject to the law of the Member State in which they are established and exempts online intermediaries from liability for the content they manage, if they fulfil certain conditions[23] (European Parliamentary Research Service, 2021).

Further to building upon the e-Commerce Directive, the DSA, and its twin piece of legislation which deals with market competition rules, the Digital Markets Act (DMA), are directly applicable as regulations, as is the intention with the AIA, which is also a proposal for a regulation. In that regard, all these documents will have binding legal force in each EU member state. In contrast, a directive outlines certain outcomes that must be achieved but each member state is free to decide how the directive is transposed into national law. The legal designation of regulations further emphasizes the significance and importance of these new rules for the Commission and the EU institutions and their wider digital regulatory agendas.

The AIA follows the Commission's "White Paper on Artificial Intelligence: A European Approach to Excellence and Trust,"[24] which was published in February 2020 as a follow up to its "European Approach to Artificial Intelligence" strategy document from April 2018.[25] The White Paper can be seen as an effort to understand via a wide consultative process whether there was a need firstly, to update regulatory checks on AI systems across sectors as they currently exist, and, if so, how they might best fit into a formal, binding legislative

framework. The White Paper proposed several ideas and structures which informed the AIA.[26] At the center of the White Paper, and as is consistent with the OECD and earlier Principles, and the framings of the DSA, DMA, and the GDPR, are concepts of safety, fairness, transparency, and data protection.

A closer look at the Digital Services Act

The liability exemptions in the e-Commerce Directive have been maintained in the DSA, where the firms that provide "transmission" (Digital Service Act, December 2020, p. 45) of information in a communication network ("mere conduit," "caching") and storage of information ("hosting" but without looking at what is being hosted) escape liability. It therefore behooves the platform to prove in individual cases that it had no "actual knowledge" of illegal content in its pages or that it has acted "without undue delay" to remove such content or block access to it.

The Commission also proposes a new set of obligations for firms of different sizes, with firms below a certain threshold considered micro-, small- and medium-sized firms and thus exempt from new transparency and disclosure obligations. Firms are defined as "very large online platforms" as those "online platforms which provide their services to a number of average monthly active recipients of the service in the Union equal to or higher than 45 million" (Digital Service Act, December 2020, p. 60)—equivalent to 10 percent of the Union's population.

The transparency obligations (Articles 26–33) for very large platforms sets this proposal apart from the GDPR. It may have been drafted with competition considerations in mind given that the GDPR reinforced, and in some cases strengthened, the market positions of large firms like Google, while smaller firms lost market share due to the disproportionate burden of costs to revenue experienced in meeting the regulation's demands (Peukert et al., 2021). The thresholds suggest that the Commission has reviewed some of the unforeseen impacts of the GDPR, though such remediation might be better met with true proportionality rather than thresholds. Nevertheless, in the draft versions at least, the 10 percent threshold as mentioned above applies to all firms that supply services to EU businesses and consumers whether they are based in Europe or not. In this aspect, the DSA will apply on an extraterritorial basis as is the case with the GDPR.

As regards content moderation, the DSA introduces several specific obligations which include, for example, the requirement that companies already carrying out content moderation must submit annual reports to authorities outlining their approach. There are further provisions regarding the takedown of illegal content with obligations to report why it was taken down (or wasn't) with EU citizens having the right to challenge and platforms required to respond within a specified time frame. The rights of appeal are explicitly outlined. These new protections for consumers have been welcomed by rights groups and consumer associations (Access Now, 2020; AlgorithmWatch, 2020).

Further obligations to strengthen consumer protection in the DSA concern advertising transparency. Article 24 obliges platforms to provide the following information: "(a) that the information displayed is an advertisement; (b) the natural or legal person on whose behalf the advertisement is displayed; (c) meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed" (Digital Services

Act, December 2020, p. 58). But these have been widely criticized as not going far enough. Rather, they are regarded as simply reinforcing the status-quo because many platforms "already allow users to see some basic information about ads and ad targeting and have created ad databases" (Blankert & Jaursch, 2021).

European Digital Rights, an influential pan-European body, whilst complimentary of the EU's attempts to provide further transparency around advertising, suggests that in this area the DSA "completely fails to address the problems inherent in the toxic ad tech business. Without any limitations to the micro-targeted online manipulation through ad tech (and with a strong ePrivacy Regulation nowhere to be seen), the constant surveillance of people across the internet for the purpose of online advertising remains the norm" (EDRI, December 2020).

A contentious section of the DSA proposal concerns the need to report on risks. For "significant systemic risks," platforms are required to identify, present a plan to mitigate, and then to report back on the efficacy of their risk mitigation measures. Reporting of anything that raises risks of dissemination of illegal content is not controversial; instead, it will be determining the risk of "any negative effects for the exercise of fundamental rights to respect for private and family life, freedom of expression and information, prohibition of discrimination and rights of the child" (Digital Services Act, December 2020, p. 61) that leaves plenty of room for dispute. On a positive note, this obligation and clause, as it has been worded, clearly shows the intent by the Commission to protect all rights as per the European Charter for Fundamental Rights (ECFR). What this means in practice, and what it implies for firms remains an active debate at the time of publication in advance of the proposal being passed into law. The DSA is not expected to be implemented until 2023.

Access to data, auditing, and enforceability

The emphasis on transparency is a very welcome part of the approach to the DSA by the Commission. For example, Article 31 on "data access and scrutiny" marks a significant and major progressive step in allowing authorities to scrutinize how platforms target, moderate, and recommend content or services to their users. Under this Article, vetted researchers will be able to apply to access platform data for the purpose of "conducting research that contributes to the identification and understanding of systemic risks" (Digital Services Act, December 2020, p. 63), including potential negative effects on fundamental rights or civic discourse.

For data access to work effectively, the U.S.-based Brookings Institute believes that the Commission:

> could set up a centralized process that enables secure data access to researchers without this capacity. The United States has effectively implemented this through systems like the Census Bureau's Statistical Research Data Centers and the Coleridge Initiative's Administrative Data Research Facility.
>
> (Engler, January 2021)

The empowering of newly created independent bodies—Digital Services Coordinators (DSCs)—at the Member State level to carry out onsite inspections is a positive step, but DSCs cannot be effective unless they are fully resourced and willing and able to use the

new powers that this act proposes. The same goes for the use of independent auditors who will undertake mandatory audits of the large platforms and that have "proven expertise in the area of risk management," as well as "technical competence" (Digital Services Act preamble, December 2020, p. 34) to audit algorithms. These requirements may indicate the need for new higher education degree programs to train enough appropriate staff for AI developing organizations, their auditors, and regulators alike. It may be sufficient to expand existing systems-engineering courses in computer science departments, but there may be a need to also to create specialist programming courses specifically for auditing and certification, rather than full-scale development. Software development, like any creative act, displays wide degrees in variation of talent but it should not be necessary to compete for the top creative talent to find those well-suited to assessing the claims being generated by the DSA.

Producing adequate competence in these areas is an achievable goal. The period of transition both before the DSA becomes law but also in the formative years of the legislation will be crucial, and lobbying efforts to dilute the effect of new laws will be significant. It is important that the DSCs are established and stabilized quickly so that commerce is not negatively impacted with too much uncertainty, and so also that there is a constructive dialogue between the DSCs and market participants as a means of allowing experiments to be tried without poor first attempts getting established as permanent precedent. Moreover, regulatory capture by agencies and individuals with vested interests is also a threat, and one that can hopefully be addressed through the same measures and standards of transparency being applied within these tools of governance as they are intended to apply without.

The Artificial Intelligence Regulation, or AI Act (AIA)

The EU deserves credit for being the first among international political organizations or states to propose a regulatory framework for updating AI systems: an immense task given the cross-cutting nature of such systems.[27] Furthermore, the EU's insistence on basing such a framework on European values and fundamental rights can hopefully again set a high global bar. In terms of international law, it would be a very significant and positive global development if its final-form regulatory framework were to achieve for the concepts of AI transparency and trustworthiness what the GDPR has achieved for the concept of privacy.

In 2018, the Commission published a strategy for AI entitled "Artificial Intelligence for Europe."[28] It is a broad and wide-ranging document laying out the EU approach towards AI over a series of steps which includes allaying public fears, updating the EU safety framework to address liability gaps and legal uncertainty, as well as announcing investments into training schemes and grants to AI-based start-ups. An overarching message was that the "EU will not leave people behind" and language deployed throughout mirrored that in the GDPR—privacy, trust, values, and ethics. The plan proposed close EU cooperation across four key areas: increasing investment, making more data available, fostering talent, and ensuring trust.

The Commission's "The White Paper: On Artificial Intelligence—A European Approach to Excellence and Trust ("the White Paper") went a step further when it proposed policy options to "enable a trustworthy and secure development of AI in Europe in full respect of the values and rights of EU citizens" (EU AI White Paper, February 2020, p. 3). The extensive

consultation that followed the White Paper showed that a large number of stakeholders "were largely supportive of regulatory intervention to address the challenges and concerns raised by the increasing use of AI" (Explanatory Memorandum, Proposal for a Regulation on a European Approach for Artificial Intelligence, April 2021, p. 1) and together with explicit requests from the European Parliament and the European Council led to the formal regulatory proposal, now referred to as the AI Act (AIA).

The AIA was announced as part of a broader package, "A European Approach to Excellence in AI," targeted to strengthen and foster Europe's potential to compete globally. Therefore, while our focus here is on the regulatory proposal itself, it is useful to understand the larger context and the accompanying coordinated plan on AI (Coordinated Plan on Artificial Intelligence 2021 Review, 2021) which details the strategy for fighting for Europe's competitiveness in AI: "Through the Digital Europe and Horizon Europe programmes, the Commission plans to invest €1 billion per year in AI. It will mobilise additional investments from the private sector and the Member States in order to reach an annual investment volume of €20 billion over the course of this decade. And the newly adopted Recovery and Resilience Facility makes €134 billion available for digital. This will be a game-changer, allowing Europe to amplify its ambitions and become a global leader in developing cutting-edge, trustworthy AI."[29] This corresponds to roughly €65 billion investment volume annually by 2025.[30]

The AIA regulatory proposal is part of a continuum of actions that started in 2017 with the European Parliament's Resolution on Civil Law Rules on Robotics and AI[31] and entailed several other key milestones[32] prior to the proposal at hand. It is addressed to AI use cases that pose a high risk to people's health, safety, or fundamental rights. All other uses of AI are explicitly *not* addressed in the new Act, with the understanding that they are already regulated by standard existing frameworks. The new regulations would apply to all providers and deployers placing on the market or putting into service high-risk AI systems in the European Union, regardless of the origin of the providing entity. In this way, the proposal seeks to level the playing field for EU and non-EU players and has mechanisms to influence far beyond its immediate scope (extra-territorial reach).

We now turn to discuss a few concepts of the AIA which appear to us to be solid and actionable. These concepts may well therefore also be the most important elements for other regions beyond the EU to consider for their own AI policy.

Clear and actionable framework for AI risk levels

The proposal suggests a risk-based approach with different rules tailored to four levels of risk: unacceptable, high, limited, and minimal risk. At the highest level of risk are systems that conflict with European values and are thus prohibited. Such a ban is a victory to all digital human rights advocates and delivers a strong message: First, do no harm. High-risk systems cover a variety of applications where foreseeable risks to health, safety, or fundamental rights demand specific care and scrutiny. According to the Commission's impact assessment, roughly five to fifteen percent of all AI systems would fall into this such high-risk category.[33] Limited risk systems are those that interact with natural persons and therefore require specific transparency measures to maintain continued human agency and to avoid deceptive uses. The remaining AI systems—that is to say seventy to ninety-five percent—fall within the lowest risk category and are not relevant for the AIA because "they

represent little or no risk for citizens' rights or safety"[34] and therefore for which the AIA introduces no new rules.

Proportionality

The AIA makes an attempt to have the requirements rightly sized in relation to the potential risks, and to regulate only what is necessary. In the AIA, the Commission presently claims to address proportionality primarily via the just-discussed risk-based approach and varied requirements depending on the system risk level. The majority of AI systems in the market would face only basic transparency requirements as mandatory if any. Nevertheless, here too true proportionality might be preferable to a threshold-based approach.

Accountability of AI supply chain (i.e., providers and deployers, not the end-users)

Another exceedingly important characteristic of the proposal is how it creates a ground for significant improvements in the supply chain transparency and accountability. No end-user can realistically be expected to take responsibility for evaluating the trustworthiness of complex technology products such as AI products. To do so, one would need a good level of transparency to the workings of the system and the technical skills necessary for meaningful evaluation. From this perspective, the Commission's choice to focus on the accountability of providers, developers, and deployers seems sensible, even if it may have led to some compromises on the end-user transparency obligations. This provider–deployer dualism is also important taking into consideration that sixty percent of organizations report "purchased software or systems ready for use" as their sourcing strategy for AI.[35]

The AIA does not suggest mechanisms that allow individual persons to submit complaints about their concerns and harm caused by AI. This has itself raised concerns in some. However, the choice seems logical considering that proper evaluation of system conformity would require much more information and technical evaluation skills than what will be available for end-users.

The solution the AIA proposes is the following: Providers are required to set up a post-market monitoring system for actively and systematically collecting, documenting, and analyzing data provided by deployers or collected otherwise on the performance of high-risk AI systems on the market. Deployers of such systems are obligated to monitor and report potential situations presenting risks. To support this mechanism's function, it seems likely that providers and deployers will implement feedback channels or contact points also for the end-users. In addition, similar feedback channels should be expected from national market surveillance authorities to support their role in identifying potential.

Meaningful documentation requirements aligned with engineering best practices

The transparency documentation requirements are as follows: Risk management system, Data and data governance, Technical documentation, Record-keeping, and Transparency and provision of information to deployers. Such documentation is in the interest of

developers for their own record keeping and the future maintenance of the product or system. While historically some companies were fooled into expensive maintenance contracts by having been persuaded "not to waste money" on also purchasing the source code from their software suppliers, there can really be no excuse for contemporary corporate boards not to know the importance of documenting the security and capacities of their software, including AI systems.

Contextual transparency reporting to AI end-users

While the main focus of the proposal is in setting specific requirements for high-risk AI systems, it is also worth mentioning what is laid down regarding transparency obligations of systems that interact with natural persons. In a short article (52) the AIA addresses what has become a major challenge with informing practices as specified in the GDPR (privacy policies): they are out of context. The requirement of the AIA is focused on the actual use context. It requires that an end-user is made aware of interacting with an AI system. This may well mean that industry standards around labeling AI products will finally start to emerge as providers begin to mark their end-user interfaces accordingly. Moreover, the AIA requires the deployers of emotion intelligence, biometric categorization, and deep fake systems to inform natural persons of their exposure to such AI systems.

Ideally, the AIA might become a new vanguard for transparency more generally. Again, taking proportionality into account, companies and other organizations may choose to publicly expose not only the minimal amount of transparency required by the law (e.g., whether the system deploys AI) but also other aspects of their transparency documentation. This should probably be done in a hierarchical way so that ordinary end-users are not overwhelmed by complexity, nor are small companies required to maintain multiple different types of documentation (which would almost certainly soon fall out of synchronization). But where providers are comfortable exposing the capacity to "drill down" into the same documentation used for regulatory and other purposes, they may find that they facilitate trust in or engagement with those systems. Some public authorities have already started to implement such transparency via public AI registers, as is also recommended by the Commission in the coordinated plan for AI.[36]

With its AIA proposal, the European Commission has shown a way to manage AI-related risks to health, safety, fundamental rights, and even social stability in a way that has all the means to incentivize industry to take appropriate action. This is of fundamental importance, offering an opportunity towards governance efficiency in regulating technologies of which influences, and impacts will be significant, but are already substantial though perhaps under-recognized.

Conclusions

The goal of this chapter has been to provide a comprehensive overview of the current situation, understandings, views, values, and regulatory approaches toward AI being deployed and under construction within the European Union with a view towards wider transnational

governance of AI systems. Given the Union's acknowledged leadership in this area, we feel this review should be of significant global value. However, we should not overlook digital regulatory advances being made not only in individual American states like California, well known for their software industry, but also more widely, including the U.S. state of Illinois, which showed leadership on facial recognition, and Asian nations such as China, India, South Korea, and Singapore. China, in particular, has also just at the time of this writing released an enormous slew of AI regulations that are in many ways in keeping with or even extending from the principles described here, as the Chinese government too seeks not only legitimacy with its own population, but also power to control its domestic software corporations.

While our review represents a comprehensive overview of the current state of the art, we must acknowledge that regulation, society, technology, and indeed the global ecological and therefore political-economic context are all highly dynamic. Nevertheless, some things that are unlikely to change, such as the critical role of regulation to stability and well-being, and of the state in enforcing such regulation. AI and other disruptive technologies must be regulated, and such regulation must be coordinated and enforced by the state. This is neither a radical nor an overly ambitious viewpoint as some (e.g., large private sector corporations) may argue. One of the primary roles of the state—arguably its very definition—is to establish and ensure quality of life and security for its citizens and residents. Clear, enforceable, and sound regulation and governance enforcement mechanisms are key to this.

As we discussed above, one highly salient but not necessarily unique aspect of AI and its regulation is the importance of global, or at least transnational, approaches and coordination. This becomes increasingly relevant when it comes to issues such as data sovereignty, social influences, and taxation. While it is unlikely and perhaps undesirable to see complete international harmonization of AI regulations at the governmental level, this chapter has highlighted how, via a process known as the Brussels effect, a well-coordinated bloc of nations with shared interests may both directly and indirectly, influence regulation internationally. If companies want to build and use AI solutions in the EU, they must comply with EU regulations, and, in many cases, the changes to their business processes may well permeate throughout their other markets. This is not only because these changes are costly, but also because they are widely perceived as providing real, positive value, including economically. It should be said that similar effects can also be observed for Beijing and Washington, although some of the Washington impacts may have been overlooked or taken for granted. But China's efforts to control what is and is not sold and operated within its national borders has also driven considerable developer effort and corporate concern in non-Chinese corporations, which has drawn the attention of legislators worldwide.

While the first transnational bespoke AI regulation is still at least a year from completion, it is of course not possible to fully predict what effect it will have globally. Yet based on previous experience with other fundamental rights-based standards and regulations, it is certainly possible to speculate that the forthcoming EU law may well have a large net positive impact for global society. More generally, regulatory activity in many regions, including the GDPR, is becoming both better informed and better understood. As such, there is good hope that it can be also better coordinated for common good, even if it remains not only varied, but necessarily driven by varied regional regulatory needs.

Notes

1. The European Commission published the Proposal for a Regulation of the European Parliament and of the Council—Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts on April 21, 2021.
2. Agreed in May 2019 by 42 states and adopted by the G20 in 2019. See Artificial Intelligence, OECD Principles on AI, available at https://www.oecd.org/going-digital/ai/principles/.
3. Annex 1 includes the following techniques and approaches:

   "(a) Machine learning approaches including supervised, unsupervised and reinforcement learning, using a wider variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines (symbolic) reasoning and expert systems; and (c) Statistical approaches, Bayesian estimation, search and optimization methods"

   (AIA Annex 1, 2021, p. 1).

4. Charlotte Stix provides a detailed overview and analysis of the EU Commission's AIA proposal in this manual under *Section 9: International Politics and AI Governance*.
5. See Artificial Intelligence, OECD Principles on AI, available at https://www.oecd.org/going-digital/ai/principles/.
6. See G20 AI Principles, available at https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf.
7. See GPAI: Global Partnership on Artificial Intelligence, available at https://gpai.ai/.
8. See Questions and Answers on the Digital Single Market Strategy, available at https://ec.europa.eu/commission/presscorner/detail/en/MEMO_15_4920.
9. See "A Union that strives for more. My agenda for Europe," available at https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf.
10. See https://europa.eu/european-union/about-eu/eu-in-brief_en.
11. Whilst much of this is derived from Europe's own history, this includes the imposition of strong antitrust regulation on Germany by the allies following WWII, and led by the United States (Wu, 2018).
12. The harmonization of standards is one of the drivers towards ensuring the integrity of the Single Market. For example, inconsistent product standards can hinder cross-border trade. The Single Market guarantees the free movement of goods and ensure seamless trade between EU member states.
13. QMV requires the support of 55 percent of member states representing the minimum of 65 percent of the EU's population.
14. See Single European Act (SEA), available at https://www.europarl.europa.eu/about-parliament/en/in-the-past/the-parliament-and-the-treaties/single-european-act.
15. See Internal Market, Industry, Entrepreneurship and SMEs, available at https://ec.europa.eu/growth/singlemarket_en#:~:text=The%20EU%20Single%20Market%20accounts,the%20protection%20of%20the%20environment.
16. Bradford (2020, p. 24).
17. EU regulatory capacity has been developed since the adoption of the Single European Act (SEA) which launched the ambitious agenda to complete the Single Market by 1992. To implement this goal, EU member states vested the EU institutions with powers to formulate and enforce regulations.

18. Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act) See https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en.
19. Proposal for a Regulation on a European Approach for Artificial Intelligence. See https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.
20. See European Convention on Human Rights, available at https://www.echr.coe.int/documents/convention_eng.pdf.
21. See "This is Why the Government Should Never Control the Internet," available at https://www.washingtonpost.com/posteverything/wp/2014/07/14/this-is-why-the-government-should-never-control-the-internet/.
22. See Directive 2000/31/EC of the European Parliament and of the Council of June 8, 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.
23. An example is the Safe Harbour Privacy Principles which were developed between 1998 and 2000 and were designed to prevent private organizations within the European Union or United States which store customer data from accidentally disclosing or losing personal information. The seven principles from 2000 are: notice, choice, onward transfer, security, data integrity, access, and enforcement (see Wikipedia, https://en.wikipedia.org/wiki/International_Safe_Harbor_Privacy_Principles).
24. See White Paper on Artificial Intelligence—A European Approach to Excellence and Trust, available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
25. See Artificial Intelligence for Europe, available at https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.
26. The White Paper was just one of several exercises that informed the AIA; any formal proposal from the Commission must be preceded by a formal impact assessment. To the end the Commission published a study supporting the impact assessment of the AIA, available at https://digital-strategy.ec.europa.eu/en/library/study-supporting-impact-assessment-ai-regulation.
27. Much of this section is taken from a working paper by two of the co-authors of this chapter; see Haataja & Bryson, "Reflections on the EU's AIA and how we could make it even better," 2021, *in preparation*.
28. See Artificial Intelligence for Europe, available at https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.
29. A European Approach to Artificial Intelligence.
30. Impact assessment accompanying the AIA, p. 70.
31. European Parliament resolution of February 16, 2017, with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).
32. For example, a report, "Ethics Guidelines for Trustworthy Artificial Intelligence," by EU AI HLEG and European Parliament resolution of October 20, 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)).
33. AIA Impact Assessment, p. 71.
34. European Commission Press Release, available at https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682.

35. European Commission, Ipsos Survey, European Enterprise Survey on the Use of Technologies based on Artificial Intelligence, 2020, p. 53.
36. Coordinated Plan on Artificial Intelligence 2021 Review by the European Commission.

Bibliography

Access Now (2020, December 15). *DSA: European Commission delivers on the first step toward systemic rules for online platforms* [press release]. https://www.accessnow.org/dsa-systemic-rules-for-online-platforms/.

AlgorithmWatch. (2020). *Automating Society Report 2020*. https://algorithmwatch.org/en/automating-society-2020/.

Anderson, David, & Murphy, Cian C. (2011). The Charter of Fundamental Rights: History and Prospects in Post-Lisbon Europe. March 31. EUI Working Papers Law 2011/08. https://ssrn.com/abstract=1894386.

Beaumeir, G., Kalomeni, K., Campbell-Verduyn, M., Lenglet, M., Natile, S., Papin, M., Rodima-Taylor, D., Silve, A., & Zhang, F. (2020). Global regulations for a digital economy: Between new and old challenges. *Global Policy, London School of Economics and Political Science 11* (4), 515–522.

Blankert, A., & Jaursch, J. (2021, January 14). What the European DSA and DMA proposals mean for online platforms. *The Brookings Institute*. https://www.brookings.edu/techstream/what-the-european-dsa-and-dma-proposals-mean-for-online-platforms/.

Bradford, Anu. (2012). The Brussels Effect. *Northwestern University Law Review 107* (1). Columbia Law and Economics Working Paper No. 533. https://ssrn.com/abstract=2770634.

Bradford, A. (2014). Exporting standards: The externalization of the EU's regulatory power via markets. *International Review of Law and Economics 42* , 158–173. http://dx.doi.org/10.1016/j.irle.2014.09.004.

Bradford, A. (2020). *The Brussels Effect: How the European Union rules the world*. Oxford University Press.

Breton, Thierry [@ThierryBreton]. (2021, January 10). A dogma has collapsed. Platforms no longer hide their responsibility by arguing that they merely provide hosting services. Capitol Hill is the 9/11 of social media. *Op-ed @POLTICO* [Tweet]. Twitter. https://twitter.com/ThierryBreton/status/1348340670856093696.

Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science 29* (2), 130–136.

Bryson, J. J. (2019). The past decade and future of AI's impact on society. In *Towards a new enlightenment? A transcendent decade*. BBVA Open Mind Series. https://www.bbvaopenmind.com/en/books/towards-a-new-enlightenment-a-transcendent-decade/.

Bryson, J. J. (2020, June 17). Regulating AI as a pervasive technology: My response to the EU AI Whitepaper Consultation [Blog Post]. https://joanna-bryson.blogspot.com/2020/06/regulating-ai-as-pervasive-technology.html.

Bryson, J. J., & Wyatt, J. (1997, April). Artificial Intelligence [Blog Post]. https://www.cs.bath.ac.uk/~jjb/web/whatisai.html.

Brundage, M., & Bryson, J. J. (2016). Smart policies for artificial intelligence. *arXiv preprint arXiv:1608.08196*.

Crootof, Rebecca, & Ard, B. J. (2020, December 18). Structuring techlaw. *Harvard Journal of Law & Technology*. http://dx.doi.org/10.2139/ssrn.3664124.

Deeks, Ashley. (2020, April 27). Introduction to the Symposium: How will Artificial Intelligence affect international law? *AJIL Unbound 114* , 138. Virginia Public Law and Legal Theory Research Paper No. 2020-38. https://ssrn.com/abstract=3586988.

Echikson, W. (2017, April 13). *Limited liability for the net? The future of Europe's e-commerce directive*. CEPS. https://www.ceps.eu/ceps-publications/limited-liability-net-future-europes-e-commerce-directive/.

EDRI. (2020, December 18). The EU's attempt to regulate big tech: What it brings and what is missing. https://edri.org/our-work/eu-attempt-to-regulate-big-tech/.

Elliott, E., Regens, J. L., & Seldon, B. J. (1995). Exploring variation in public support for environmental protection. *Social Science Quarterly 76* (1), 41–52. http://www.jstor.org/stable/44072588.

Engler, Alex. (2021, January 15). Platform data access is a lynchpin of the EU's Digital Services Act [Blog Post]. https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/.

Erdelyi, Olivia Johanna, & Goldsmith, Judy, (2018, February 2). Regulating artificial intelligence: Proposal for a global solution. 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, New Orleans. https://ssrn.com/abstract=3263992.

European Commission. (2017, January 10). *Communication from the Commission to the European Parliament and the Council: Exchanging and protecting personal data in a globalised world* [communication]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2017%3A7%3AFIN.

European Commission. (2020). Proposal for a regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. https://ec.europa.eu/info/sites/info/files/proposal_for_a_regulation_on_a_single_market_for_digital_services.pdf.

European Commission. (2020). On artificial intelligence—A European approach to excellence and trust [White paper]. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Commission. (2020, September 16). State of the Union Address 2020 [speech]. https://ec.europa.eu/info/sites/default/files/soteu_2020_en.pdf.

European Commission. (2020, December 3). European Democracy Action Plan: making EU democracies stronger [press release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_20_2250.

European Commission. (2020, December 15). Digital Markets Act: Ensuring fair and open digital markets [press release]. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2349.

European Commission. (2020, December 15). Digital Services Act—Questions and answers [press release]. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348.

European Commission. (2021). Coordinated plan on Artificial Intelligence 2021 review. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council: Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

European Commission. (2021). Study supporting the impact assessment of the AU regulation. https://digital-strategy.ec.europa.eu/en/library/study-supporting-impact-assessment-ai-regulation.

European Commission. (2021, March 26). Speech by Commissioner Reynders at American Chamber of Commerce on the Digital Transatlantic Economy. *Check Against Delivery*. https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_21_1482.

EU Fundamental Rights Agency. (2020). *Getting the future right: Artificial intelligence and fundamental rights*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf.

EU High-Level Expert Group AI. (2019). AI HLEG Update. https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=30466#:~:text=%E2%97%A6%20Definition%20of%20A.I.%3A%20%E2%96%AA,collected%20structured%20or%20unstructured%20data%2C.

European Parliament. (2017). Report with recommendations to the Commission on the Digital Services Act: Improving the functioning of the Single Market. https://www.europarl.europa.eu/doceo/document/A-9-2020-0181_EN.html.

European Parliamentary Research Service. (2021, March). Digital Services Act [briefing]. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689357/EPRS_BRI(2021)689357_EN.pdf.

Evans, B. (2020, July 23). Regulating technology. *Benedict Evans* [blog]. https://www.ben-evans.com/benedictevans/2020/7/23/regulating-technology.

Fioretti, J. (2016, June 6). German privacy regulator fines three firms over U.S. data transfers. *Reuters*. https://www.reuters.com/article/us-germany-dataprotection-usa-idUSKCN0YS23H.

The Global Partnership on Artificial Intelligence. (2020, November). Global partnership on Artificial Intelligence: Term of reference. https://gpai.ai/about/gpai-terms-of-reference.pdf.

Goldthau, A., & Sitter, N. (2015). Soft power with a hard edge: EU policy tools and energy security. *Discussion Paper, Review of International Political Economy*. https://www.tandfonline.com/doi/full/10.1080/09692290.2015.1008547?journalCode=rrip20#abstract.

Guasch, J. L., & Hahn, R. W. (1999). The Costs and Benefits of Regulation: Implications for developing countries. *Policy Research Working Papers, World Bank Research Observer*, 137–138. https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-1773.

G20. (2019). G20 AI Principles. https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf.

Harari, Y. N. (2021, February 26). Lessons from a year of Covid. *The Financial Times*. https://www.ft.com/content/f1b30f2c-84aa-4595-84f2-7816796d6841.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1* , 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Lauchlan, S. (2020, January 23). WEF 2020—There is no 'Plan B' if time runs out on a global deal on digital services taxation, admits OECD. diginomica. https://diginomica.com/wef-2020-there-no-plan-b-if-time-runs-out-global-deal-digital-services-taxation-admits-oecd.

Lorenz, Philippe. (2020, September 14). AI Governance through political fora and standards developing organizations—Mapping the actors relevant to AI governance. *Mapping AI Governance Fora*. http://ssrn.com/abstract=3731354.

Lynskey, Orla. (2020, August 15). Extraterritorial impact in data protection law through an EU law lens. Brexit Institute Working Paper Series—No 8/2020. https://ssrn.com/abstract=3674413.

Mărcuț, M. (2017). *Crystalizing the EU digital policy: Exploration into the Digital Single Market*. Springer.

Miguel, J. C., & Casado, M. Á. (2016). GAFAnomy (Google, Amazon, Facebook and Apple): The big four and the b-ecosystem. In M. Gómez-Uranga, J. M. Zabala-Iturriagagoitia, & J.

Barrutia (Eds.), *Dynamics of big internet industry groups and future trends: A view from epigenetic economics* (pp. 127–148). Cham: Springer International Publishing.

OECD. (2019, May 22). Artificial intelligence—OECD Principles on AI. https://www.oecd.org/going-digital/ai/principles/.

OECD Legal Instruments. (2019). Recommendation of the Council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#dates.

O'Reilly, T. (2017). *WTF? What's the future and why it's up to us*. New York: Random House.

Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2021). Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science* (forthcoming). https://ssrn.com/abstract=3560392.

Pew Research Center. (2018, August 17). How Americans and Western Europeans compare on 4 key social and political issues. https://www.pewresearch.org/fact-tank/2018/08/17/how-americans-and-western-europeans-compare-on-4-key-social-and-political-issues/.

Pfeiffer, S., & Carr, R. (2021). Error 404—trust not found: A European survey on digital (dis)trust. Munich Security Conference, Munich Security Brief 2, March. https://doi.org/10.47342/REFQ1817.

Preusse, Verena, Purnhagen, Kai Peter, & Dries, Liesbeth. (2020, September 3). The "Brussels Effect" in action: Do the EU's standards travel abroad? A study on control systems for geographical indications in the EU and Thailand. Bayreuth Working Paper Series Food Law 02/2020. https://ssrn.com/abstract=3685915.

Ram, N., & Gray, D. (2020). Mass surveillance in the age of COVID-19. *Journal of Law and the Biosciences 7* (1), lsaa023. https://doi.org/10.1093/jlb/lsaa023.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society 7* . https://journals.sagepub.com/doi/full/10.1177/2053951720942541.

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2019, December 18). What's next for AI ethics, policy, and governance? A global overview. https://doi.org/10.1145/3375627.3375804.

Schwartz, Paul M. (2019). Global data privacy: The EU way. *New York University Law Review 94* (771). https://ssrn.com/abstract=346554.

Schwartz, Paul M., & Peifer, Karl-Nikolaus. (2017). Transatlantic data privacy. *Georgetown Law Journal 106* (115). UC Berkeley Public Law Research Paper. https://ssrn.com/abstract=3066971.

Scott, J. (2014). Extraterritoriality and territorial extension in EU Law. *American Journal of Comparative Law 62* (1). https://ssrn.com/abstract=2276433.

Shalal, A., Nienaber, M., & Thomas, L. (2021, February 26). U.S. drops "safe harbor" demand, raising hopes for global tax deal. *Reuters*. https://www.reuters.com/article/us-g20-usa-oecd-idUSKBN2AQ2E6.

Silver, L. (2018, August 17). *How Americans and Western Europeans compare on 4 key social and political issues*. Pew Research Center. https://www.pewresearch.org/fact-tank/2018/08/17/how-americans-and-western-europeans-compare-on-4-key-social-and-political-issues/.

Singel, R. (2008, April 8). EU tells search engines to stop creating tracking databases. *Wired*. https://www.wired.com/2008/04/eu-tells-search/.

Stix, Charlotte. (2021, February 19). Actionable principles for artificial intelligence policy: Three pathways. *Sci Eng Ethics 27*(15). https://ssrn.com/abstract=3792370.

Technology Council Committee on Technology, N. S. a. (2016). Preparing for the future of artificial intelligence. Technical report, Executive Office of the US President. https://obam

awhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Veale, Michael, & Zuiderveen Borgesius, Frederik. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International 22* (4). https://ssrn.com/abstract=3896852.

Vestager, M., & Borrell, J. (2021, March 10). Why Europe's digital decade matters. *Project Syndicate*. https://www.project-syndicate.org/commentary/europe-digital-decade-by-margrethe-vestager-and-josep-borrell-2021-03?barrier=accesspaylog.

Venkataramakrishnan, S. (2020, April 27). GDPR accused of being toothless because of lack of resources. *Financial Times*. https://www.ft.com/content/a915ae62-034e-4b13-b787-4b0ac2aaff7e.

Whittlestone, J., Nyrup, R., Alexandova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society, 195–200. https://dl.acm.org/doi/10.1145/3306618.3314289.

Wu, T. (2018). The curse of bigness. Antitrust in the new gilded age. Columbia Global Reports.

CHAPTER 14

STANDING UP A REGULATORY ECOSYSTEM FOR GOVERNING AI DECISION-MAKING

Principles and Components

VALERIE M. HUDSON

Artificial Intelligence Decision-Making (AIDM) refers to automated systems which make decisions formerly made by human beings, on the basis of information provided to the system exogenously or compiled by it. These systems can be quite trivial, with little import for society or human security, and with fully human-specified algorithms at work, such as a system of "decision-making" to activate an emergency sprinkler when a heat sensor registers a certain temperature. Over time, however, the significance of the systems for society and human security has risen, for example in the case of automated consumer loan decisions or of self-driving vehicles—or of automated weapons systems. In other cases, the ability of the AIDM to mine data sources without human intervention and "train itself" on data it compiles in order to create and implement algorithms that are not completely under the control of or even not completely understood by human beings, is also a cause of increasing social and personal concern. Some AIDM merely assists human decisionmakers, but in other cases the decisionmaker is the AIDM system itself. At this point in time, the horse has already left the proverbial barn in terms of governing AIDM, and governments have been left to play catch up with technology firms that have faced little to no regulation to date.[1]

That situation is changing, especially with new moves by the European Union to develop oversight capabilities concerning AIDM.[2] Indeed, there are now several guiding documents being developed and disseminated on the topic, including the 2016 EU General Data Protection Regulation (GDPR),[3] the 2019 OECD Principles on Artificial Intelligence,[4] and more recently the April 2021 European Commission's new Proposal for a Europe-wide Artificial Intelligence Act,[5] and the May 2021 UK Guidance on an Ethics, Transparency, and Accountability Framework for Automate Decisionmaking.[6] In this chapter, I approach the question of the governance of AI decision-making (G/AIDM) in a holistic fashion, asking

what would the overall "ecosystem" of G/AIDM have as component elements in order to render such governance both robust and sustainable over time? How could checks and balances be woven into the very structure of that ecosystem, so that it remains functional, sustainable, and adaptive? While this chapter is focused primarily on the United States and its existing governance structure, the principles and functions discussed know no national boundaries. In addition, this chapter is interested in the use of AIDM by both the private and the public sectors.

Principles Guiding Regulatory Ecosystem Creation

Before a regulatory system can be stood up, its foundations must be laid in guiding principles. In other work,[7] I have asserted that the principles upon which G/AIDM is built must center the rights of humans vis-à-vis AIDM systems that affect their lives. As the IEEE—the premier U.S. professional association for computer scientists and electrical engineers—states, "Autonomous and Intelligent Systems (A/IS) should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control."[8] These human rights, at a minimum, should include the right to *know*, the right to *appeal*, and the right to *litigate*.

The right to know

Every human has the right to know when they are engaging with an AI system. Beyond simple notification that they are encountering an AI system, individuals should have unfettered access to a standardized identification label with the contact information for the party with a fiduciary obligation for the performance of the system.

The proposed EU guidelines state that "AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system."[9] Identification is also important because it is anticipated that G/AIDM law will indicate domains in which AIDM systems cannot be legally used, such as in decisions whether to use lethal force. Furthermore, without the human right to know, the rights of appeal and of harm-based litigation become moot. Alex Engler of the Brookings Institution notes, "There's immediate value to a broad requirement of AI-disclosure: reducing fraud, preserving human compassion, educating the public, and improving political discourse."[10]

This demands both central record-keeping and situational notification.[11] Before an AIDM system can be deployed, it should be registered with the government. That is, the government must be notified it exists, who the creator and the deployer are, and archive the code itself or determine that the code has been adequately archived. The IEEE concurs: "Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS."[12] This is a function that ideally would have been created years ago; there is no such existing central governmental record-keeping capability in the U.S. as of this writing.

A standard identity tagging system will be an important element of G/AIDM, and no AIDM system would receive permission for deployment until that tag is in place and accessible to humans with which the system will be interacting. Such an identity tag would be made available to humans interacting purposefully with the system, and would at a minimum include the identification number under which the AIDM system is registered with the government. With more passive interaction, such as facial recognition software applied to CCTV footage, the registration number would be made available upon request. This might well become a Miranda-like right in the case of law enforcement AIDM.

And having the tag simply available may not be enough. It may be necessary for certain AIDM systems to identify up front that they are not human; that is, situational notification will be necessary. Steven Brykman provides an interesting example of how it is no longer a straightforward matter to decide whether one is interacting with a human being or an AI system:

> Everyone I know has gotten the "IRS" call claiming they owe the government money. But nobody falls for it because the dude making the call is clearly not a native speaker and his patter is ridiculous. But imagine if a bot were making the call! Suddenly, the English is perfect, and the details are legitimately convincing. Hell, AI can even be used to replicate real people's voices! Even voices of people we know—by grabbing snippets of their voice from videos on Facebook. Talk about fake news! The AI could even pull actual information about the call recipient from big data like their social media accounts, say, and then incorporate that info into the conversation to help "prove" its identity. And it would all seem totally natural. Just like how magicians pretend to read the minds of audience members by using information the theater already asked them for when they bought the tickets.[13]

California has moved forward with its "Blade Runner" bill that mandates such bots identify themselves ("I am a chat bot") if they are making a sales pitch or trying to influence an election, though social media companies at present cannot be held liable for these bots operating on their platforms.[14] (The new EU AI regulations[15] actually prohibit "practices that have a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities.") In addition, the state of Illinois has passed an Artificial Intelligence Video Interview Act, which requires that job applicants be informed in writing that an AI system is being used as part of a video interview, along with an explanation of what information the AI system will be examining, and how the system evaluates that information.[16] The U.S. requires a national approach to "the right to know," especially as such interactive systems are capable of gross invasions of privacy, providing data-hoovering businesses with a plethora of new data points about our personal lives.[17] Government registration before deployment, coupled with standardized identity tags and situationally mandated notifications, will be essential in preserving the human right to know they are interacting with an AIDM system.

To be noted in passing is that the right to know implies that there may also be a concomitant right to refuse to interact with a non-human system. The EU notes, "The option to decide against this [AI] interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights."[18] While that issue will not be tackled in this paper, it is one which calls for a national discussion as the regulatory ecosystem components are put in place.

The right to appeal

Every human has the right to appeal an AIDM decision to a human being, not another AIDM system. As the IEEE asserts, "Individuals should be provided a forum to make a case for extenuating circumstances that the A/IS may not appreciate—in other words, a recourse to a human appeal."[19] The right to appeal to a human being is the means by which the subordination of AIDM to humankind is effected. This is also in concert with the OECD principle insisting that government "ensure that people understand AI-based outcomes and can challenge them."[20] Article 22 of the EU's GDPR specifically asserts that, "(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her; (3) the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least *the right to obtain human intervention on the part of the controller*, to express his or her point of view and *to contest the decision*" (emphasis mine).[21] Under the UK's GDPR, for example, there is actually an Article 22 checklist that must be applied pre-deployment of any AIDM system.[22]

We would argue that the right to appeal to a human being is foundational to successfully incorporating AI into a social system. Consider that in June 2020, officials in Kentucky announced they would host an open-air meeting area near the state capitol building for citizens to speak with officials about problems with their unemployment benefits. Many had not yet been paid, and the pandemic lockdown was in its fourth month. The state-deployed algorithmic phone tree that never allowed human contact no matter what combination of options you pushed was experienced as completely, even maddeningly, unhelpful. When word leaked out that you had the chance to speak to a real human being—in person even!—people traveled all night from all over Kentucky to be there when it opened at 9 am. According to news reports,[23] the first person arrived at 3:00 am, and by 10:30 am, police had to tell anyone who just arrived that they would not be able to see an official that day because thousands were already in line, and just seeing those who had arrived before that time would take eight full hours.

So eight full hours they stood, spaced according to social distancing requirements, with their masks and their papers, in the Kentucky sun. Those that couldn't be seen came back the next day, and even then, some had to be turned away again due to the crowds. Some had borrowed money to make it to the state capitol and had to sleep in the surrounding parks because they had no money for a hotel room. What drove them there? We argue there is a profound human need to appeal to other human beings in cases of distress.

Furthermore, this human being must be an official representative of the entity making the decision, empowered to change the decision made by the AIDM system. Though humans are by no means foolproof, human beings can more readily see when an algorithm has veered from its intended purposes in terms of outcomes generated.[24] Successful appeals should catalyze a renewed audit of the algorithm in question. Consider that in January 2020, a facial recognition algorithm used by the police department in Detroit, Michigan, led officers to arrest Robert Julian-Borchak Williams for larceny.[25] Surveillance camera footage from the store in question was fed into an algorithm provided by DataWorks Plus, and the algorithm attempted to match the footage to drivers' license photos, ultimately

deciding Williams might be the culprit. He was arrested and cuffed in front of his family, and brought to the station. What happened next is worth reflection:

> In Mr. Williams's recollection, after he held the surveillance video still next to his face, the two detectives leaned back in their chairs and looked at one another. One detective, seeming chagrined, said to his partner: "I guess the computer got it wrong."

Note that the humans involved recognized the algorithm was dead wrong—and they recognized it in a nanosecond. Indeed, the human mind excels at the ability to hold in mind both the big picture and the minute details that comprise it. In addition, the human stakes are (hopefully) meaningful to other human beings; false arrest holds significant importance to humans. Appeal to a human being is thus a necessary complement to AIDM systems. The need for appeal was also clearly shown in the infamous COMPAS case[26] where a biased recidivism prediction algorithm resulted in differential treatment of offenders based on race, but this bias was only demonstrated by independent third-party analysis of the data.

The right to appeal, and the logically related capability of overturning, an AIDM decision requires *explicability* of the decision made. The IEEE asserts, "Common sense in the A/IS and an ability to explain its logical reasoning must be required"[27] and the EU concurs that "the discussion on the so called 'right to explanation' when automated decision-making occurs is important to address. Even though it is not yet guaranteed in Europe, future jurisprudence or Member State laws could grant individuals the right to ask for an explanation when a solely automated decision (e.g., refusal of an online credit application or e-recruiting practices) is being made about them that has legal or other significant effects. Such a right could provide a mechanism to increase the transparency and accountability of A/IS, and should therefore be seriously considered."[28] This right to appeal of course includes the right to see and correct one's personal information, and to have explained how that data informs the decision taken. Current UK guidance mandates that, "Process owners need to introduce simple ways for the impacted person(s) to request human intervention or challenge a decision. When automated or algorithmic systems assist a decision made by an accountable officer, you should be able to explain how the system reached that decision or suggested decision in plain English."[29]

Thus right of appeal thus necessitates the creation of *standards of explicability* of the AIDM system's methods for reaching a decision. As the EU Guidelines note, "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process . . . . In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available."[30] The IEEE adds, "laws could grant individuals the right to ask for an explanation when a solely automated decision (e.g., refusal of an online credit application or e-recruiting practices) is being made about them that has legal or other significant effects."[31] What counts as a sufficient explanation is an important conceptual undertaking for any nascent regulatory system. We have seen in other areas of human endeavor, such as human subject research, the necessity of setting standards for informed consent, which include mandated elements such as explanations of purposes, risks, and benefits of that research, using language that is easy to understand, and with contact information if the human subject has questions or concerns. A similar effort to develop standards for AIDM will be needed.

The right to appeal to a human empowered to overturn or modify a decision made by an AIDM also entails the building of such capability to overturn/modify. That is, the construction of an appeal interface and the training of individuals to take on the role of "appellate judges" in the process will be important tasks. Human bureaucrats, as we all know, are fully capable of erroneous judgment, too. As one civil servant articulated, "if your case is decided by one of my colleagues, they'll follow the spirit of the rule and you'll get a reasonable decision; if your case is decided by the person next door, they'll stomp their foot and insist on zealously following the letter of the rules ('my hands are tied,' 'I can't do anything,' or the one that makes my blood boil, 'it would be unfair to treat this case especially') . . . Those people aren't even malicious, just brain-dead. You just have to pray and hope you never find yourself under their power."[32] Is it possible that a combination of AIDM and appeal to human beings might correct the worst excesses of each?

The right to litigate

Every human has the right to **litigate** the harm caused by AI applications. Legal liability should rest, I argue, with the vendor of the AIDM system who has sold the product to the entity whose deployment of the AI system caused harm. It is vendors that are most intimately knowledgeable about the system, and most capable of preventing or reducing harm their systems might cause. Now, this does not preclude the idea that purchasing entities that could be found legally negligent in their purchase from the vendor, for example, having failed to check for due diligence on the part of the vendor with regard to procurement, in terms of appropriate pre-deployment testing and auditing, and thus through such negligence could also be held partially liable.

While we consider the timeline of due diligence more specifically in the next section, fiduciary principles must be established.[33] That is, the purchasing entity must assume the vendor of the AIDM system is providing it with an agent (the AIDM system) that will act according to the preferences of the principal (the purchasing entity). At a minimum, then, the vendor owes the entity that purchases the system due diligence to the entity's preferences, including the preference to mitigate risk for which it might be held liable. To whom else is the vendor of an AIDM system responsible? The vendor also owes the human beings who will interact with its AIDM system a responsibility, and it also entails a responsibility to the national society in which the AIDM is deployed. Note that it is entirely possible that these responsibilities may conflict, which is another justification for a regulatory system that can specify which fiduciary responsibilities trump which others under law.

The vendor of an AIDM system should, as a first step towards fulfilling its responsibilities, ensure the system is lawful in every aspect from data collection to decision implementation. As noted previously, governments will define areas and tasks for which AIDM may not be used; government will also put forward regulations, such as the requirement for identity tags and pre-deployment registration, that must be followed.

As a second step, the vendor must ensure that its AIDM system actually performs the function for which the purchasing entity procured it. Performance standards will invariably be both generalizable and non-generalizable across systems. That is, one can imagine performance standards based on generalizable issues such as accuracy, or match with human decision-makers' decisions. But one can also imagine performance standards

that are very much task-specific. There is a role for issue-oriented government entities in helping to set such standards, as the U.S. Food and Drug Administration (FDA) has done in establishing a Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD).[34] These types of frameworks may or may not need to be supplemented by the purchasing entity to suit its purposes.

Vendors will surely come under regulations for due diligence to human beings who are affected by the AIDM, as well as to the broader national society. There is a growing literature on the possible harmful effects of AIDM on both individuals and collectives. Specific, personal harm to particular individuals, as well as class-based or characteristic-based harm, must both be litigable. In the case of the latter, the plaintiff may well be the national government, which might bring suit for violation of federal law on, say, non-discrimination grounds. But to be litigable, the concept of a "tort" must be expanded under law to include not only conventionally understood harms (a self-driving car runs over a pedestrian, for example), but the types of harm that characteristically result from the use of an algorithm to make a decision about a human being or the use of an algorithm to target an individual for an intervention (such as an influence attempt). In addition, government itself may be the offending party, and blanket immunity for the government under common tort law would be a nightmare for citizens and requires amendment. Infantino and Wang comment that, "many features of algorithms might constitute a challenge for tort law systems," and we agree.[35] They elucidate:

> Algorithms' virtual ubiquity and massive use in both the private and the public sectors give them the potential to cause not only singular, specific accidents but also widespread and recurring mass injuries. Many algorithm-related activities span across borders, therefore triggering complex issues about jurisdiction (and applicable law). What is more, in case of an accident, a number of algorithmic features—from algorithms' hermetic language and technological opaqueness, to their complex authorhood, their possible interaction with the environment and self-learning capacities, their interdependence with other systems and/or with physical devices—raise fundamental problems for making the factual determinations necessary to allocate responsibility and for determining what went wrong, why things went wrong, how and when they did, and what/who caused the injury . . . Algorithmic torts pose distinctive challenges to potential claimants, who might face daunting problems in identifying the source of the wrong that befell them, selecting the appropriate defendants and venue for litigation, gathering evidence, searching for litigation funding, overcoming the threshold of governments' immunity, and so on and so forth.

There is no doubt that the concept of due diligence will require third parties outside of the vendor and the purchasing entity which intends to deploy the system. In-house certification of AIDM systems will not be sufficient to protect both parties (vendor/creator and purchaser/deployer) from lawsuits. These independent third parties—which we discuss more fully in Part III—must exist for purposes of pre-deployment testing, deployment auditing, and expert witness in case of litigation. Given the proprietary nature of source code, these third parties will enable the safeguarding of such code while allowing for presumably unfettered analysis. These third parties should be empowered to offer recommendations to government oversight entities for revision or even for forced de-deployment for enforcement purposes. In addition, as AIDM systems are retired and/or vendors go out of business, third-party archiving will also be essential for "zombie" systems that continue to be used.

Courtroom standards with regard to litigation will also need to be promulgated for the right to litigate to be meaningful. For example, though the bill died in Congress, something like H.R. 4368 (2019) will be necessary.[36] Called the "Justice in Forensic Algorithms Act," the legislation would have ensured that defendants have access to source code and other information necessary to exercise their due process rights when algorithms are used to analyze evidence in their case as the *State (Wisconsin) v. Loomis* case highlights.[37] In the case a defendant has been harmed by an AIDM, the defendant must also have to right to force the vendor to reveal the source code to third-party expert witnesses who can analyze how the harm was produced by the AIDM system. Indeed, the AI Now Institute in Australia has gone further than this, recommending that "all public agencies that use AIDM systems should require vendors to waive any trade secrecy or other legal claim that might inhibit algorithmic accountability, including the ability to explain a decision or audit its validity."[38] This will be a critical legal element in any viable regulatory ecosystem governing AIDM.

I would argue that this standard apply both to public agencies and private deployers of AIDM. Consider the case of "Gregg," a young man whose predicament was detailed in *The New York Times* recently.[39] He gambled with an app called Sky Bet, and the company deploying Sky Bet had amassed an entire dossier on Gregg, even his banking and mortgage records. Sky Bet had contracts with various companies to hoover up information on Gregg, including one called Iovation, which "provided a spreadsheet with nearly 19,000 fields of data, including identification numbers for devices that Gregg used to make deposits to his gambling account and network information about where they were made from." But it was what came after the massive data collection, as fraught as that is, that is even more concerning. *The Times* recounted how software appeared to offer suggestions to lure back Gregg after he stopped gambling in late 2018. In the data profile that listed Gregg as a customer to "win back," there were codes noting he was receptive to gambling promotions that featured Las Vegas. Having made more than 2,500 deposits on Sky Bet, he was listed as a "'high value' customer." Here we have moved from data collection to algorithmic decision-making concerning an individual human being, with an aim to altering that individual's behavior against his own personal interest towards the financial interest of Sky Bet. Notice the interventions promoted by this AIDM:

> After [Gregg] stopped gambling, Sky Bet's data-profiling software labeled him a customer to "win back." He received emails like one promoting a chance to win more than $40,000 by playing slots, after marketing software flagged that he was likely to open them. A predictive model even estimated how much he would be worth if he started gambling again: about $1,500.

The potential for harm here is enormous, as real-life cases illuminate: "A high-flying engineer killed himself hours after an online casino 'groomed' him with £400 in cash bonuses. Gambling addict Chris Bruney, 25, lost a total of £119,000 in the five days before his death, but instead of shutting his account, Winner.co.uk plied him with cash bonuses and free bets."[40]

The legal strategy to demonstrate harm to "Gregg" and others is possible only if legal counsel has access to the code that translates Gregg's data into recommendation for actions that Sky Bet should take to keep Gregg betting. This right to access code must be legislated, and should cover those who have been harmed either by government or private deployment of AIDM systems.

Finally, as noted previously, the right to litigate assumes the national government must have executive capacity to force non-deployment, revision, or de-deployment of AIDM

systems.[41] It will most likely take the form of a government AIDM oversight agency, as we will outline in Part III. This new entity of the federal government would also have the power to assess penalties for non-compliance. While this would be new territory for the U.S., European nations have moved in this direction already; for example, the UK has an Information Commissioner's Office (ICO) to enforce the public's "information rights"[42] and has already issued helpful guidance about topics such as implementing the rights to explanation and appeal.[43]

Due Diligence: Pre-Deployment, Deployment, and Retirement

In this section, we turn from the principles underlying the creation of a new regulatory system for AIDM to a timeline of what needs to happen when in terms of performing due diligence with respect to AIDM systems; that is, when an AIDM system is developed, through deployment, and eventual retirement. This exercise will help us to further understand what function-based components will be a necessary part of that regulatory system, which system we will begin to flesh out in Part III.

Pre-Deployment

Measures to be taken pre-deployment by vendors or others developing AIDM systems would include:

1. Training of developers in the law surrounding deployment of an AIDM system, as well as the compliance requirements.
2. Registration of an AIDM system with the federal agency having oversight over AIDM systems. (If the software incorporates extant registered systems, those must be included/referenced in the registration.) Such registration includes the archiving of code with the agency, as well as the creation of an identity tag for the system which must be deployed with the system and be accessible to those who will interact with it. If a system is updated or changed in any way, the identity tag must reflect which version is being encountered by the humans interacting with it.
3. Initial testing of the AIDM for bias, discrimination, harm, risk, and so forth.[44] This testing must be both in-house and with an accredited third-party AIDM testing body. The results of this testing, along with the test data used, must be archived by the vendor/developer. No identity tag should be issued by the government without evidence of this dual testing demonstrating benignity. The IEEE provides additional recommendations: "Automated systems should generate audit trails recording the facts and law supporting decisions and such systems should be amenable to third-party verification to show that the trails reflect what the system in fact did. Audit trails should include a comprehensive history of decisions made in a case, including the identity of individuals who recorded the facts and their assessment of those facts.

Audit trails should detail the rules applied in every mini-decision made by the system. Providers of A/IS, or providers of solutions or services that substantially incorporate such systems, should make available statistically sound evaluation protocols through which they measure, quality assure, and substantiate their claims of performance, for example, relying where available on protocols and standards developed by the National Institute of Standards and Technology (NIST) or other standard-setting bodies."[45]

4. A system for continuous monitoring of impacts, as well as a system to assess the effect of software updates/changes on the system's impacts (including the archiving of all versions of the AIDM), must be in place before deployment.
5. Insurance for the harms produced by the AIDM system must be purchased; a harms payment fund, akin to the National Vaccine Injury Compensation Fund, should be established by public entities desiring either to build or to deploy AIDM systems.[46]
6. If an AIDM will be deployed by a government, a pre-deployment public comment period should be undertaken. While there is more citizen choice in interacting with private entities' deployment of AIDM systems, there may be little to no choice for citizens interacting with governmental AIDM, and concerns should be able to be raised by citizens before deployment.

Deployment

1. If a government entity is not the developer but is buying a system from a vendor/developer, AI Now rightly argues that the purchase contract "should ensure the contract includes language requiring the vendor to guarantee the product or service is compliant with the relevant antidiscrimination laws. Inclusion of such clauses will ensure that government agencies have legal standing to have the system fixed, and that vendors too have liability if AIDM use produces discriminatory outcomes."[47] This type of contractual assurance for procurement purposes would be useful for private deployers, as well; some legal experts have also called for the creation of corporate board AI oversight committees to mitigate risk, as well, to demonstrate due diligence.[48] Typical procurement criteria, such as pricing, must not overshadow compliance criteria.[49] For far too long, argue Brauneis and Goodman (2018), "governments simply did not have many records concerning the creation and implementation of algorithms, either because those records were never generated or because they were generated by contractors and never provided to the governmental clients. These include records about model design choices, data selection, factor weighting, and validation designs. At an even more basic level, most governments did not have any record of what problems the models were supposed to address, and what the metrics of success were."[50] That must change, and it will only change through regulation.
2. A user interface that satisfied the "right to know" and "the right to appeal" must be implemented at time of deployment. This would alert human beings interacting with a system that an AIDM system is in use, it would provide the identity tag, typically digital in nature, and it would provide an interface for appeal to a human decision-maker.
3. Human appeals referees would need to be trained and deployed concomitant with deployment of the AIDM. The ability to periodically audit the work of these referees

should also be in place. A mechanism for public complaint via à vis the referees themselves should also be established in this regard.

4. Periodic testing of the AIDM system should occur to see if additional real-world inputs alter the judgments concerning harm, bias, and discrimination. This testing would be performed by a neutral third-party assessment organization, and would generate additional audit trails for analysis. Ditto for significant updates to the AIDM system's code, which should trigger independent audits as well. Such significant updates would also need to be logged with the federal agency tasked with AIDM oversight and registration, and an updated identity tag attached to the system.
5. Should periodic testing results show harm, discrimination, or bias, the federal agency tasked with AIDM oversight would be notified by the independent auditors, and this federal agency would have the power to force de-deployment of the system until the problems are rectified.

Retirement of an AIDM System

1. A vendor or developer wishing to retire an AIDM system would be required to notify the federal agency tasked with AIDM oversight about this intention, and the effective date of retirement would be noted by the agency. This agency should retain all registration records and archived code with testing data for a period of seven to 10 years.
2. The vendor or developer must certify that the AIDM system is not in use anywhere by anyone before such retirement can be made effective. Capabilities to provide evidence that is the case must be developed.
3. The federal agency must retain the ability to examine how archived code functions over that seven to 10 year period. That is, it must retain legacy hardware and software systems necessary to run the archived programs for at least that period of time.
4. A general statute of limitations for harms related to AIDM should be coincident with the archiving period of seven to 10 years.

Components of a Functioning Regulatory Ecosystem

On the basis of our discussion of principles as well as the timeline of AIDM assurance, let us recap and summarize what institutions and what procedures/functionalities need to exist within a robust regulatory ecosystem for AIDM oversight and harm mitigation. Figure 14.1 attempts that summary:

The starting point for the regulatory system depicted in Figure 14.1 rests in the human rights to know, to appeal, and to litigate their experience with AIDM systems, whether the systems are deployed by the government or by private entities. To ensure these rights, there must be effort to create structures and processes across that create a comprehensive ecosystem, touching actors in law, government, technology, business, and consumer advocacy. Unfortunately, regulation of AIDM lags behind its deployment. Hopefully with a holistic

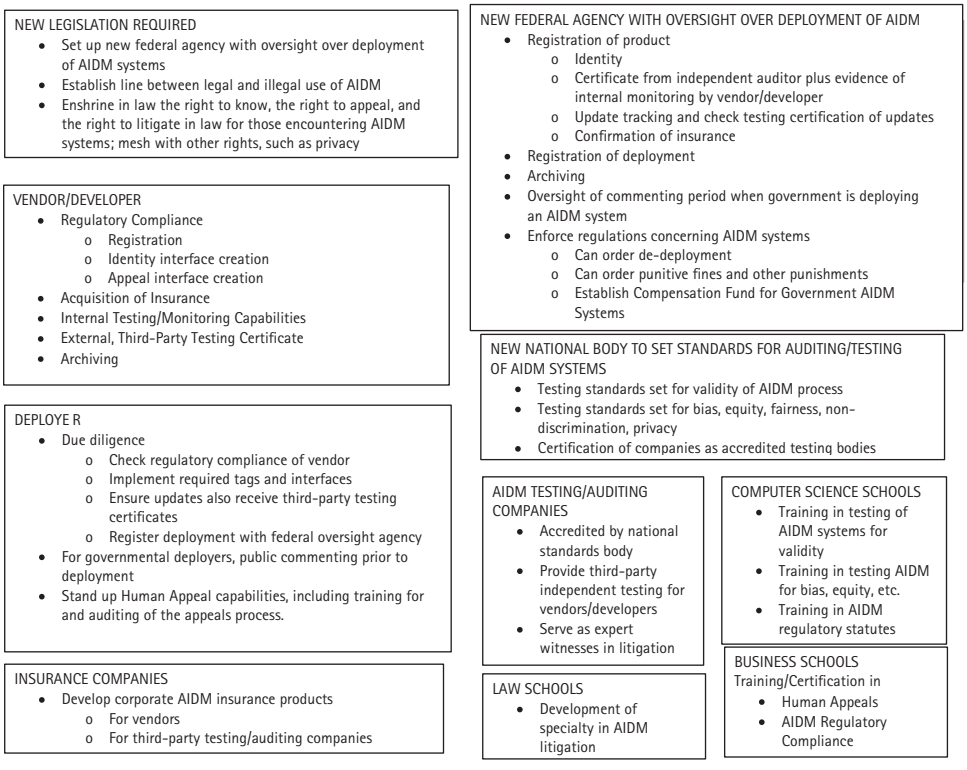
NEW LEGISLATION REQUIRED
Set up new federal agency with oversight over deployment of AIDM systems
Establish line between legal and illegal use of AIDM
Enshrine in law the right to know, the right to appeal, and the right to litigate in law for those encountering AIDM systems; mesh with other rights, such as privacy
VENDOR/DEVELOPER
Regulatory Compliance
Registration
Identity interface creation
Appeal interface creation
Acquisition of Insurance
Internal Testing/Monitoring Capabilities
External, Third-Party Testing Certificate
Archiving
DEPLOYE R
Due diligence
Check regulatory compliance of vendor
Implement required tags and interfaces
Ensure updates also receive third-party testing certificates
Register deployment with federal oversight agency
For governmental deployers, public commenting prior to deployment
Stand up Human Appeal capabilities, including training for and auditing of the appeals process.
INSURANCE COMPANIES
Develop corporate AIDM insurance products
For vendors
For third-party testing/auditing companies
NEW FEDERAL AGENCY WITH OVERSIGHT OVER DEPLOYMENT OF AIDM
Registration of product
Identity
Certificate from independent auditor plus evidence of internal monitoring by vendor/developer
Update tracking and check testing certification of updates
Confirmation of insurance
Registration of deployment
Archiving
Oversight of commenting period when government is deploying an AIDM system
Enforce regulations concerning AIDM systems
Can order de-deployment
Can order punitive fines and other punishments
Establish Compensation Fund for Government AIDM Systems
NEW NATIONAL BODY TO SET STANDARDS FOR AUDITING/TESTING OF AIDM SYSTEMS
Testing standards set for validity of AIDM process
Testing standards set for bias, equity, fairness, non-discrimination, privacy
Certification of companies as accredited testing bodies
AIDM TESTING/AUDITING COMPANIES
Accredited by national standards body
Provide third-party independent testing for vendors/developers
Serve as expert witnesses in litigation
LAW SCHOOLS
Development of specialty in AIDM litigation
COMPUTER SCIENCE SCHOOLS
Training in testing of AIDM systems for validity
Training in testing AIDM for bias, equity, etc.
Training in AIDM regulatory statutes
BUSINESS SCHOOLS
Training/Certification in
Human Appeals
AIDM Regulatory Compliance

FIGURE 14.1 Components of a Functioning Regulatory Ecosystem

vision of what is required to uphold these right, important institutions and capabilities can be stood up. The government plays a pivotal role, and I argue that what is urgently needed *now* in the United States is legislation to create a new federal agency with oversight over AIDM deployment, accompanied by legislation that sets out the lines of legal and illegal use of AIDM and enshrines the rights to know/appeal/litigate in law.

This needed legislation must provide the foundation to integrate law on deployment of AIDM with other areas of law, such as privacy rights, the principle of non-discrimination, and so forth. The most important consequence of such legislation would be the standing up of a new federal agency with specific oversight over the deployment of AIDM systems. The agency could be either stand-alone, or more likely subsumed under the Federal Trade Commission's mandate. (The European Commission, by comparison, suggests the creation of a European Artificial Intelligence Board, which would have many of these same responsibilities.)[51]

This new agency of the U.S. government would have several functions. First, it would elucidate the government regulations under which AIDM is legal to deploy. It might even set out areas where deployment is legal but discouraged and disincentivized because of the potential for harm.[52] Second, it would register AIDM systems created by vendors/developers at the point where the system would be deployed by the vendor/developer, or be sold to deploying entities. After checking that the vendor/developer has performed internal and external testing of the system for validity and legality, created the requisite interfaces,

archived the code, and confirmed it has acquired insurance for the system, the agency would issue a registration identifier to be used in the required identity tag enshrining the right to know. The new agency would also keep an archive of the code. In sum, this new agency would ensure the system has undergone what the European Commission calls a "conformity assessment," which would result in a certification or the denial of such certification before the system can be sold for deployment or deployed.[53] Upon deployment of the system, either by the vendor/developer or by a separate deploying entity, the deployment would be registered with the new agency as well. If the deploying entity is part of government, the new agency would also oversee a public commenting opportunity before deployment. Importers of systems developed in other countries would have to submit the system to the agency for a conformity assessment before deployment.

The new agency would have enforcement powers; that is, it would have the power to order rapid de-deployment of an AIDM system shown to be harmful or illegal, as well as punitive fines and other forms of punishments. Updates or modifications to an AIDM system would also be subject to registration and renewed testing mandates. As noted previously, we suggest an archive period of seven to 10 years for litigation purposes. Given the new agency's registry, the continued deployment of a vintage AIDM system can readily be observed due to its identity tag. It is also possible that lawmakers may want to take the new agency with periodic "algorithmic impact assessments" in key social sectors such as lending, banking, and educational AIDM systems to look for overall changes in national social equity, bias, discrimination, etc.[54]

But the new agency by itself cannot constitute the entire regulatory ecosystem; it will depend on the creation and standardization of new capabilities within the national marketplace. One of the most pivotal of these is the creation of a national standards board for the testing of AIDM systems, both for validity testing (i.e., does it perform the function desired) and testing for other legal and societal goods, such as explicability, non-discrimination, privacy, etc. This new board will lay down standards for due diligence in testing and due diligence in purchasing such systems. Regarding testing, the creation of these standards will then allow the standards board to certify that private, independent testing/auditing companies are following best practices in making their judgments.

With regard to due diligence in purchasing tested systems, we have seen the development of new standards boards in recent years to react to changing political priorities. For example, those interested in climate change created the Sustainability Accounting Standards Board, which lays out what data needs to be tracked and what kinds of financial disclosures are necessary to ensure companies are operating with environmental sustainability in mind.[55] Similarly, the Financial Stability Board has weighed in on what they call "TCFD"—climate-related financial disclosures—that companies can use to report to their investors. An "ethical AIDM standards board" can therefore not only begin to enumerate what types of testing and auditing are necessary to ensure ethical use of AIDM, but also help specify in more detail what corporate due diligence with reference to AIDM involves in terms of monitoring and disclosure.[56]

The new board's standards will thus enable the marketplace to welcome these accredited AIDM testing/auditing companies as part of the regulatory ecosystem, and make their testing determinations part of the due diligence responsibility of vendors and purchasers. These testing/auditing companies may be new, or existing companies may

expand into this area. In addition to being hirable by vendors/developers of AIDM systems for purposes of fulfilling the testing certification requirements of the new federal agency, these companies can also serve as expert witnesses when AIDM system harm is litigated in the courtroom.

For their part, vendor/developers must not only have the capability to create AIDM systems, but must have a parallel compliance capability. That is, these vendor/developers must have internal testing capabilities, for it would not be economical to hire an accredited third-party testing company without first having assessed that the AIDM system is likely to pass muster. Vendor/developers are also under regulatory obligation to create identification and appeal interfaces as a native part of the AIDM system, acquire insurance for their system, and be capable of archiving the code for their system.

Sometimes an AIDM system will be deployed by the developer itself, but in many cases the system will be sold to a separate entity that desires to deploy it. The deployer also has regulatory obligations. The first and foremost is to ensure that the vendor/developer from which they are purchasing the system has complied with all federal and state regulations: this is the heart of due diligence on the part of the deployer, and is the foundation for asserting non-responsibility in litigation cases. If the deploying entity is part of the government, the new agency in charge of AIDM oversight should oversee a public commenting period, as previously discussed. The deployer must faithfully enable all identification and appeal interfaces, and must register deployment with the new federal agency for AIDM oversight. Last, the deployer must stand up robust human appeal capabilities, and monitor the appeals process for adequacy of response. While deployers may not be responsible for harms caused by the AIDM system if they have performed due diligence, they may be held responsible for harms caused by a wholly inadequate human appeals process.

There are some natural ancillary capabilities that must exist within the ecosystem; one good example is insurance. Existing insurance companies will need to create insurance products for vendors/developers, for third-party independent testing companies, and perhaps even for deployers with relation to the human appeals process. Given the creativity of insurance companies to adapt to changing societal conditions and innovations, we have no doubt that this can be done in a fairly expeditious fashion.

But there are other ancillary capabilities that will not be so expeditious to put into place, but which are critical to the functioning and sustainability of the system. And here is where our institutions of higher learning must be tapped. Training and certification of training will be essential elements of a system that ensures human rights to know, to appeal, and to litigate. For example, the new federal agency as part of the U.S. government will be responsible to prevent or even reverse AI system deployment where it cannot be shown that risks have been satisfactorily mitigated. But what training and standards will be required for government actors to make such decisions?

Because the training and the standards are not yet in place, the human capital that will be necessary to ensure citizen rights with reference to AI systems does not yet exist. As Brundage and Bryson (2016) note, "The most important and obvious thing that governments should do to increase their capacity to sensibly govern AI is to improve their expertise on the matter through policy changes that will support more talent and the formation of an agency to coordinate AI policy. It is widely recognized that governments are ill-equipped currently

to deal with ongoing AI developments."[57] Here we will consider at least three such requisite efforts that societies must expeditiously undertake if AI governance is to be effective in ensuring citizen rights:

1. ***Recruit and train independent algorithm auditors.*** We envision that for a healthy AI ecosystem to exist, independent and impartial auditing firms that check the claims made by tech companies will be necessary in the private sector. These AIDM auditors would need to be certified by the standards board previously discussed. The federal oversight agency for AIDM will not have the resources to perform all the audits required; this is a cost that the public should not have to pay. But there are also other parts of the regulatory ecosystem that will need trained, accredited auditors: the federal oversight agency will need such individuals, and private companies will also need them for internal compliance purposes. In other words, trained, certified AIDM system auditors will need to be hired by vendor/tech companies, by independent testing/auditing firms, and by governments. That is a tall order, and universities—probably computer science programs within universities—will be called upon by the market to provide that human capital.
2. ***Create a new track within the Human Resources (HR) field in business schools that focuses on the training of new HR Human Appeals Officers for deployed AI systems.*** The HR field already trains professionals in relevant skills, such as grievance mediation, identification of discriminatory differential treatment, legally compliant decision making with regard to hiring, and other direct effect processes, among many others. Training HR professionals to handle legally mandated human appeals processes for AI systems is a natural extension of this field of expertise. It is important for entities deploying AIDM systems to have trained, certified personnel in these appeal processes, for this is the point at which litigation against a deploying entity is likely to be successful.
3. ***Develop the capabilities of lawyers and legal professionals to litigate algorithm liability cases.*** Law schools with ambition understand that a specialty in AI and the law will become a growth field. Legal professionals will also benefit from basic education in the auditing of algorithms, since many expert witnesses in these cases will be professional algorithmic auditors. This is an opportunity for computer science programs to offer a non-specialist course in understanding how algorithms work and how one can assess if an algorithm is performing in an illegal manner or has produced individual or social harm.

We understand that mobilizing these capabilities will be a daunting task, especially for universities that have faced budget cutbacks. It may be necessary for the federal government to provide seed funding for the creation of the training and certification programs necessary to get the regulatory ball rolling. This would not be unprecedented: the government has in the past provided multi-year grants to universities (and individual scholarships to students) to develop capabilities in area studies and languages, for example, with an eye to recruiting students from these programs into intelligence agencies in the future. The rational and pro-social regulation of AIDM systems is no less important a task for the government.[58]

Conclusion

The horse is already half-way out the barn door; AIDM systems are becoming natural extensions of existing technologies in decision-making and in influencing decision-making.[59] It is time for the U.S. government to strategically move into this space and close the barn door before it becomes impossible to catch and bridle the horse. Rather than stumble into regulation piecemeal, given the immense possible individual and social harms that can come from unregulated AIDM systems, we urge a holistic and comprehensive view of what will be needed to stand up a healthy, functioning, and sustainable regulatory ecosystem for AIDM.

Such an ecosystem will include several moving parts, including founding legislation, a new federal oversight agency, a new standards board, regulations and statutes, and the development of significant new capabilities in both the private and public sectors. It will involve insurance companies, testing/auditing companies, vendors and developers, deploying entities, and universities. The founding legislation establishing a federal oversight agency is the first order of business. But the tallest order is that of manpower. Given that trained, certified human capital is the key to each of the needed capabilities in the regulatory ecosystem for AIDM, the government should seriously consider seed-funding to universities to create the programs necessary to fill our currently considerable gaps.

Only with such a working regulatory ecosystem can the fundamental rights of humans in relation to AIDM systems be maintained: the right to know, the right to appeal, and the right to litigate. If we fail to act expeditiously, it may be impossible to recapture these rights once they have been lost.

Notes

1. Zuboff, S. (2020). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.
2. European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
3. Intersoft Consulting. (2018). General Data Protection Regulation (GDPR). Retrieved on June 28, 2021, https://gdpr-info.eu/.
4. OECD. (2021, June 28). Artificial Intelligence—OECD Principles on AI. Retrieved May 6, 2021. https://www.oecd.org/going-digital/ai/principles/.
5. European Commission. (2021). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106(COD). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
6. Office for Artificial Intelligence. (2021, May 21). *Guidance: Ethics, transparency and accountability framework for automated decision-making*. UK Cabinet Office, Central Digital & Data Office, Office for Artificial Intelligence. Retrieved June 28, 2021. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making.

7. Abecassis, A., Bullock, J. B., Himmelreich, J., Hudson, V. M., Loveridge, J., & Zhang, B. (2020, June 26). *Contribution to a European agenda for AI: Improving risk management, building strong governance, accelerating education and research*. Berkman Klein Center for Internet & Society at Harvard University. https://medium.com/berkman-klein-center/contribution-to-a-european-agenda-for-ai-13593e71202f.
8. IEEE. *The IEEE global initiative on ethics of autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
9. European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
10. Engler, A. (2020). *The case for AI transparency requirements*. Center for Technology Innovation at Brookings. https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/.
11. Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law and Technology 20* (103), 103–176. https://yjolt.org/sites/default/files/20_yale_j._l._tech._103.pdf.
12. IEEE. *The IEEE global initiative on ethics of autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
13. Brykman, S. (2018, June 13). Why we desperately need an AI chatbot law. *CIO*. https://www.cio.com/article/3281375/why-we-desperately-need-an-ai-chatbot-law.html.
14. Engler, A. (2020). *The case for AI transparency requirements*. Center for Technology Innovation at Brookings. https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/.
15. European Commission. (2021). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106(COD) https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
16. Illinois Artificial Intelligence Video Interview Act, Ill. Stat. § 101-0260 (2020). http://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260.
17. Dickson, B. (2018, July 17). Why AI must disclose that it's AI. *PCMag*. https://www.pcmag.com/opinions/why-ai-must-disclose-that-its-ai.
18. European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
19. IEEE. *The IEEE global initiative on ethics of autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
20. OECD. (2021, June 28). *Artificial Intelligence—OECD Principles on AI*. Retrieved May 6, 2021. https://www.oecd.org/going-digital/ai/principles/.
21. Article 22 GDPR. (2018). *Automated individual decision-making, including profiling*. General Data Protection Regulation (EU GDPR). https://gdpr-text.com/read/article-22/.
22. Information Commissioner's Office. Rights related to automated decision-making including profiling. Retrieved June 28, 2021. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/.

23. Griffith, K. (2020, June 17). Thousands of people stand in jaw-dropping line for EIGHT HOURS to speak to officials about their unpaid unemployment benefits in Kentucky. *Daily Mail*. https://www.dailymail.co.uk/news/article-8433495/Thousands-stand-Kentucky-unemployment-line-EIGHT-HOURS.html.
24. Goodyear, S. (2020). How a U.K. student's dystopian story about an algorithm that grades students came true. *CBC Radio*. https://www.cbc.ca/radio/asithappens/as-it-happens-the-wednesday-edition-1.5692159/how-a-u-k-student-s-dystopian-story-about-an-algorithm-that-grades-students-came-true-1.5692437.
25. Hill, K. (2020, August 3). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html?action=click&module=Top%20Stories&pgtype=Homepage.
26. Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
27. IEEE. The IEEE global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
28. European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
29. Office for Artificial Intelligence. (2021, May 21). Guidance: Ethics, transparency and accountability framework for automated decision-making. UK Cabinet Office, Central Digital & Data Office, Office for Artificial Intelligence Retrieved June 28, 2021. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making.
30. European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
31. IEEE. *The IEEE global initiative on ethics of autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
32. Jordan, M. (2020, July 7). A woman without a country: Adopted at birth and deportable at 30. *The New York Times*. https://www.nytimes.com/2020/07/07/us/immigrants-adoption-ice.html#permid=107989637.
33. Mitnick, B. M. (2020). *The theory of agency: The fiduciary norm*. SSRN. https://ssrn.com/abstract=3681014.
34. U.S. Food & Drug Administration. *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)*. https://www.fda.gov/media/122535/download.
35. Infantino, M., & Wang, W. (2018). Algorithmic torts: A prospective comparative overview. *Transnational Law & Contemporary Problems 29* (1). https://ssrn.com/abstract=3225576.
36. Text of H.R.4368—116th Congress (2019-2020): Justice in Forensic Algorithms Act of 2019. (2019, October 2). https://www.congress.gov/bill/116th-congress/house-bill/4368/text.
37. Harvard Law Review. (2017). *State vs. Loomis:* Wisconsin Supreme Court requires before use of algorithmic risk assessments in sentencing. *Harv. L. Rev. 130*, 1530. https://harvardlawreview.org/2017/03/state-v-loomis/.

38. AI Now Institute, New York University. (2020). Submission to the Australian human rights & technology discussion paper. https://ainowinstitute.org/ai-now-comments-to-the-australian-human-rights-commission.pdf.
39. Satariano, A. (2021, April 1). What a gambling app knows about you. *The New York Times*. https://www.nytimes.com/2021/03/24/technology/gambling-apps-tracking-sky-bet.html.
40. Witherow, T. (2020, May 26). Tragic gambler who was 'groomed' with a bonus: how online casino plied 25-year-old with a £400 booster just hours before he took his own life. *The Daily Mail*. https://www.dailymail.co.uk/news/article-8359275/Online-casino-plied-Chris-Bruney-25-400-boost-hours-took-life.html.
41. Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now*. https://ainowinstitute.org/aiareport2018.pdf.
42. The Information Commissioner's Office. (2019). https://ico.org.uk.
43. Information Commissioner's Office. (2020). *Guidance on the AI auditing framework: Draft guidance for consultation*. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.
44. For one example of how this can be done, with sex as the biasing factor, see Smith, G. & Rustagi, I. (2021, March 31). When good algorithms go sexist: Why and how to advance AI gender equity. *Stanford Social Innovation Review*. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity.
45. IEEE. *The IEEE global initiative on ethics of autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.
46. National Vaccine Injury Compensation Program. (2021). Health Resource & Services Administration. https://www.hrsa.gov/vaccine-compensation/index.html.
47. AI Now Institute, New York University. (2020). Submission to the Australian human rights & technology discussion paper. https://ainowinstitute.org/ai-now-comments-to-the-australian-human-rights-commission.pdf.
48. Fontenot, L., & Gaedt-Sheckter, C. (2020, January 3). Fiduciary duty considerations for boards of cos. using AI. *Law 360: GibsonDunn*. https://www.gibsondunn.com/wp-content/uploads/2020/01/Fontenot-Gaedt-Sheckter-Fiduciary-Duty-Considerations-For-Boards-Of-Cos.-Using-AI-Law360-1-3-2020.pdf.
49. Mulligan, D. K., & Bamberger, K. A. (2019). Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34, 773. https://lawcat.berkeley.edu/record/1137218?ln=en; McBride, K., van Noordt, C., Misuraca, G., & Hammerschmid, G. *Towards a systematic understanding on the challenges of procuring artificial intelligence in the public sector*. Unpublished manuscript.
50. Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law and Technology 20* (103), 103–176. https://yjolt.org/sites/default/files/20_yale_j._l._tech._103.pdf.
51. European Commission. (2021). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106(COD). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

52. Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 61* (4), 66–83. https://doi.org/10.1177/0008125619862257; https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/swales41&id=1130&men_tab=srchresults.
53. European Commission. (2021). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106(COD). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
54. Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now*. https://ainowinstitute.org/aiareport2018.pdf.
55. Value Reporting Foundation. SASB Standards. (2018). https://www.sasb.org.
56. Task Force on Climate-Related Financial Disclosures. (2017). https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf.
57. Brundage, M., & Bryson, J. (2016). Smart policies for artificial intelligence. https://arxiv.org/ftp/arxiv/papers/1608/1608.08196.pdf.
58. Eidelson, B. (2021). Patterned inequality, compounding injustice, and algorithmic prediction. *The American Journal of Law and Equality (Forthcoming)*, *Harvard Public Law Working Paper* 21(14), 27 pages. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3790797.
59. Pinkstone, J. (2021, April 5). Are you part of the 'Emojification resistance'? Scientists urge people to pull faces at their phone as part of a new game that exposes the risks of 'emotion recognition technology.' *The Daily Mail*. https://www.dailymail.co.uk/sciencetech/article-9436813/Scientists-urge-public-try-new-game-risks-emotion-recognition-technology.html.

References

Abecassis, A., Bullock, J. B., Himmelreich, J., Hudson, V. M., Loveridge, J., & Zhang, B. (2020, June 26). *Contribution to a European agenda for AI: Improving risk management, building strong governance, accelerating education and research*. Berkman Klein Center for Internet & Society at Harvard University. https://medium.com/berkman-klein-center/contribution-to-a-european-agenda-for-ai-13593e71202f.

AI Now Institute, New York University. (2020). *Submission to the Australian human rights & technology discussion paper*. https://ainowinstitute.org/ai-now-comments-to-the-australian-human-rights-commission.pdf.

Article 22 GDPR. (2018). *Automated individual decision-making, including profiling*. General Data Protection Regulation (EU GDPR). https://gdpr-text.com/read/article-22/.

Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law and Technology 20* (103), 103–176. https://yjolt.org/sites/default/files/20_yale_j._l._tech._103.pdf.

Brundage, M., & Bryson, J. (2016). Smart policies for artificial intelligence. https://arxiv.org/ftp/arxiv/papers/1608/1608.08196.pdf.

Brykman, S. (2018, June 13). Why we desperately need an AI chatbot law. *CIO*. https://www.cio.com/article/3281375/why-we-desperately-need-an-ai-chatbot-law.html.

Dickson, B. (2018, July 17). Why AI must disclose that it's AI. *PCMag*. https://www.pcmag.com/opinions/why-ai-must-disclose-that-its-ai.

Eidelson, B. (2021). Patterned inequality, compounding injustice, and algorithmic prediction. *The American Journal of Law and Equality (Forthcoming), Harvard Public Law Working Paper 21* (14), 27 pages. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3790797.

Engler, A. (2020). *The case for AI transparency requirements*. Center for Technology Innovation at Brookings. https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/.

European Commission. (2020). White paper on artificial intelligence—A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Commission. (2021). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final} Brussels, 21.4.2021 COM(2021) 206 final 2021/0106(COD). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

Fontenot, L., & Gaedt-Sheckter, C. (2020, January 3). Fiduciary duty considerations for boards of cos. using AI. *Law 360: GibsonDunn*. https://www.gibsondunn.com/wp-content/uploads/2020/01/Fontenot-Gaedt-Sheckter-Fiduciary-Duty-Considerations-For-Boards-Of-Cos.-Using-AI-Law360-1-3-2020.pdf.

Goodyear, S. (2020). How a U.K. student's dystopian story about an algorithm that grades students came true. *CBC Radio*. https://www.cbc.ca/radio/asithappens/as-it-happens-the-wednesday-edition-1.5692159/how-a-u-k-student-s-dystopian-story-about-an-algorithm-that-grades-students-came-true-1.5692437.

Griffith, K. (2020, June 17). Thousands of people stand in jaw-dropping line for EIGHT HOURS to speak to officials about their unpaid unemployment benefits in Kentucky. *Daily Mail*. https://www.dailymail.co.uk/news/article-8433495/Thousands-stand-Kentucky-unemployment-line-EIGHT-HOURS.html.

Harvard Law Review. (2017). *State vs. Loomis:* Wisconsin Supreme Court requires before use of algorithmic risk assessments in sentencing. *Harv. L. Rev. 130*. 1530. https://harvardlawreview.org/2017/03/state-v-loomis/.

Hellman, D. (2020). Sex, causation, and algorithms: How equal protextion prohibits compounding prior injustice. *Wash. U.L. Rev. 98* (281). https://openscholarship.wustl.edu/law_lawreview/vol98/iss2/7.

Hill, K. (2020, August 3). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html?action=click&module=Top%20Stories&pgtype=Homepage.

IEEE. (n.d.) The IEEE global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf.

Illinois Artificial Intelligence Video Interview Act, Ill. Stat. § 101-0260 (2020). http://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260.

Infantino, M., & Wang, W. (2018). Algorithmic torts: A prospective comparative overview. *Transnational Law & Contemporary Problems 29* (1). https://ssrn.com/abstract=3225576.

Information Commissioner's Office. (2020). *Guidance on the AI auditing framework: Draft guidance for consultation*. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

Information Commissioner's Office. *Rights related to automated decision making including profiling*. Retrieved June 28, 2021, from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/.

Intersoft Consulting. (2018). *General Data Protection Regulation (GDPR)*. Retrieved on June 28, 2021, https://gdpr-info.eu/.

Jordan, M. (2020, July 7). A woman without a country: Adopted at birth and deportable at 30. *The New York Times*. https://www.nytimes.com/2020/07/07/us/immigrants-adoption-ice.html#permid=107989637.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

McBride, K., van Noordt, C., Misuraca, G., & Hammerschmid, G. *Towards a systematic understanding on the challenges of procuring artificial intelligence in the public sector*. Unpublished manuscript.

Mitnick, B. M. (2020). The theory of agency: The fiduciary norm. *SSRN*. https://ssrn.com/abstract=3681014.

National Vaccine Injury Compensation Program. (2021). Health Resource & Services Administration. https://www.hrsa.gov/vaccine-compensation/index.html.

Mulligan, D. K., & Bamberger, K. A. (2019). Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34, 773. https://lawcat.berkeley.edu/record/1137218?ln=en.

OECD. (2021, June 28). *Artificial Intelligence—OECD Principles on AI*. Retrieved May 6, 2021. https://www.oecd.org/going-digital/ai/principles/.

Office for Artificial Intelligence. (2021, May 21). *Guidance: Ethics, transparency and accountability framework for automated decision-making*. UK Cabinet Office, Central Digital & Data Office, Office for Artificial Intelligence. Retrieved June 28, 2021.https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making.

Pinkstone, J. (2021, April 5). Are you part of the 'Emojification resistance'? Scientists urge people to pull faces at their phone as part of a new game that exposes the risks of 'emotion recognition technology.' *The Daily Mail*. https://www.dailymail.co.uk/sciencetech/article-9436813/Scientists-urge-public-try-new-game-risks-emotion-recognition-technology.html.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now*. https://ainowinstitute.org/aiareport2018.pdf.

Satariano, A. (2021, April 1). What a gambling app knows about you. *The New York Times*. https://www.nytimes.com/2021/03/24/technology/gambling-apps-tracking-sky-bet.html.

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 61* (4), 66–83. https://doi.org/10.1177/0008125619862257.

Smith, G., & Rustagi, I. (2021, March 31). When good algorithms go sexist: Why and how to advance AI gender equity. *Stanford Social Innovation Review*. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity.

Sourdin, Tania. (2018). Judge v. Robot? Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal 41* (4), 1114–1133.

Task Force on Climate-Related Financial Disclosures. (2017). https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf.

Text of H.R.4368—116th Congress (2019–2020): Justice in Forensic Algorithms Act of 2019. (2019, October 2). https://www.congress.gov/bill/116th-congress/house-bill/4368/text.

The Information Commissioner's Office. (2019). https://ico.org.uk.

U.S. Food & Drug Administration. (n.d.) *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)*. https://www.fda.gov/media/122535/download.

Value Reporting Foundation. *SASB Standards*. (2018). https://www.sasb.org.

Witherow, T. (2020, May 26). Tragic gambler who was "groomed" with a bonus: how online casino plied 25-year-old with a £400 booster just hours before he took his own life. *The Daily Mail*. https://www.dailymail.co.uk/news/article-8359275/Online-casino-plied-Chris-Bruney-25-400-boost-hours-took-life.html.

Zuboff, S. (2020). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.

CHAPTER 15

LEGAL ELEMENTS OF AN AI REGULATORY PERMIT PROGRAM

BRIAN WM. HIGGINS

Introduction

In the modern U.S. regulatory era, permission-type governance approaches have traditionally not been lawmakers' first choice for regulating new technologies. Consider, for example, the U.S. national regulatory permit programs for new drugs/medical devices and stationary air pollution systems, both of which evolved after decades of alternative regulatory approaches. Morgan (2017) suggests one reason for this is that lawmakers, being aware of the problems of imposing rigid rules too early, may adopt a wait-and-see approach for new technologies.[1] Bell and Ibbetson (2012) suggest something more structural: the inherent slow and incremental *ex post* nature of the legislative process, the exception being when "a particular event or moral panic creates political pressure to which the legislature needs to provide a rapid response."[2] Moses (2017) frames the issue using the so-called Collingridge dilemma, in which regulators must choose between early regulation, when there are many unknowns about a technology's trajectory, risks, and benefits, and later regulation, when technological frames become less flexible.[3] Whether for those or other reasons, a formal permission governance strategy is also not among the levers of governance currently being considered in the U.S. for regulating artificial intelligence (AI) technologies (Fischer et al., 2021).[4]

Regulatory permit programs, like the ones mentioned above, have been around for decades, yielding important benefits for society (DeMuth, 2012). Most share a common legal approach: government authorities, acting as agents for the public, intercede in the life cycle of newly developed products and services to review and clear them before they enter the market. Once cleared, specific operating permits are issued, which are written licenses or warrants from a person in authority empowering the grantee to do some act not forbidden by law, but not allowable without such authority (Black, 1990). Enforcement is achieved by laying down terms and conditions in the permit and regularly monitoring for compliance.

It is the argument of this chapter that a national regulatory permit program should be given proper consideration as part of prudential AI governance. As seen, individual, AI system-specific permits are a reasonable legal response to the far-reaching risks posed by AI, and they satisfy lawmakers' search for "safe, responsible, and democratic means" for AI's further development.[5] At the very least, a national regulatory permit program, with a full economy-wide reach and *ex ante* risk mitigation approach, may protect the public's interests and personal rights[6] better than some of the alternatives, including industry self-governance, soft-law expectation setting, permit-less command-and-control structures, and civil tort litigation.

The purpose of this chapter is to provide a framework for thinking about how to permit AI technologies by describing the essential legal elements lawmakers could include in an AI permit program.[7] The next section of this chapter focuses precisely on these elements, including a means for addressing applicability (answering the contentious question: who and what should be permitted to operate?), how authorities might review permit applications containing highly complex and nuanced technical information, and what terms and conditions an AI system operating permit might contain. As well, attention is given to administrative and judicial enforcement elements, both of which are crucial to ensuring responsible parties remain accountable to those most likely impacted by their systems. Drawing inspiration from existing regulatory permit approaches, as well as pending U.S. federal legislation concerning AI, exemplary statutory language is provided to frame the discussion.

It is acknowledged that a national regulatory permit program for AI technologies with a sufficiently wide net will burden the AI industry, at least because new regulations have a direct economic cost, but also because of possible deployment delays caused by government reviews (which is not a trivial matter, given the relatively rapid pace at which AI technologies develop). Moreover, a permit program will be expensive and difficult to get up and running and is not expected to remove all risk associated with AI technologies. Additionally, a formal review and permit program of the kind envisioned here may become less effective as the industry moves beyond "narrow" AI technologies (those that can be functionally well-defined and operate in a specific vertical) to more advanced systems yet to be discovered or that autonomously generate other AI systems (a legal grey area if there ever was one). Furthermore, as Biber and Ruhl (2014, 2015) caution, permit programs create barriers to markets and thus enlarge the advantage deep-pocket incumbents already have in their respective markets. There may also be unintended consequences of government intervention, with possible wide-ranging repercussions (DeMuth, 2012).

Even so, the benefits of a national permit program for AI, at least in comparison to the alternatives, are not insignificant. The sections that follow address some of those benefits and other relevant issues. In particular, "The Politics of Permits: Government Intrusion, Federalism, and the Status Quo" section briefly reflects on the reasons a permit program could succeed despite foreseeable criticisms, juxtaposing its anticipated benefits against existing tort litigation and soft-law approaches. The section entitled "Europe's Proposed AI Regulations, a U.S. Program for Medical Devices, and the 'General' Permission Regulatory Approach" compares the legal elements proposed here to the recent European regulatory approach announced in April 2021, and to a proposal by the U.S. Food and Drug Administration (FDA) for approving AI-powered medical devices.

Legal Elements of an AI Permit Program

> Possible legislative language: No covered person may place an AI system in the stream of commerce or commence operation of an AI system without first obtaining a permit or certificate as described by this Act.

The *sine qua non* of the AI permit program is a new legislative act containing prohibitory language of the kind suggested above, and which authorizes one or more federal agencies to establish a formal technology and risk review program along with the discretion to issue specific (individualized) permits to "covered persons."[8] As indicated, the law could prohibit covered persons from placing an AI system in commerce or operating an AI system unless they have been issued a permit.[9] Key elements of the law, or its implementing regulations, are covered in the sub-sections below. They include: (a) identifying who are covered persons; (b) identifying the information they need to submit to a reviewing agency for it to fairly assess an AI system; (c) identifying standards and criteria for properly evaluating those systems; (d) establishing permit issuance procedures, including public participation; (e) establishing monitoring and recordkeeping requirements to ensure compliance; and (f) providing enforcement mechanisms, including statutory penalties for violations and judicial review.

Applicability determination element

> Possible legislative language (definition): COVERED PERSON—is any person conducting business or providing services in the [United States] who either (i) deploys a Listed AI System for their benefit or on behalf of another, (ii) has annual gross revenues or income of at least [$5 million] and either, (a) deploys or uses an AI technique to process data of at least [250,000] consumers or (b) derives over 50 percent of gross revenue or income from the sale of products or services embodying an AI technology, or (iii) is substantially owned, operated, or controlled by a person, partnership, institution, or corporation that meets the requirements under clause (ii).

Establishing who and what is covered by the AI permit program is critical. Regulate too many AI systems, and innovation may suffer. Too few, and systems that ought to be regulated may be allowed to operate unchecked by government authorities. The permit/approval programs run by the FDA and the U.S. Environmental Protection Agency (EPA) exempt from most permit requirements a tranche of low-risk and low-impact devices or systems.[10] In the example statutory language above, this concept is similarly followed. The suggested approach comprises automatically regulating any recognized high risk systems ("Listed AI Systems"), as well as larger entities that, due to the sheer volume of data they process or revenue earned from AI technologies, may pose a potential risk and thus should be scrutinized even if their AI systems might not be categorically high-risk (Figure 15.1).

Analogous to the way the EPA regulates listed "criteria" and hazardous air pollutants,[11] which are known to cause acute and chronic health effects, a list of AI systems or technologies could be drawn up from among those whose functions are known to have

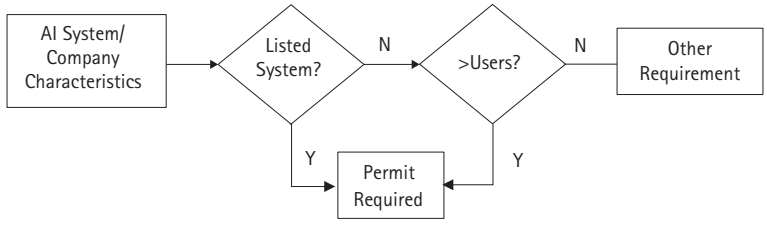

FIGURE 15.1 A general process for identifying a "covered person" based on its AI system and/or individual or company characteristics

high potential risk of interfering with or infringing the public's interests or personal rights. The list, developed initially by lawmakers and then supplemented by regulators on a regular frequency, might include systems and technologies that (1) surveil humans and their behaviors or activities, (2) collect user personal data, (3) deliver healthcare decisions or services, (4) classify individuals according to a set of features for purposes of allocating public resources to those individuals, (5) generate synthetic media or content either autonomously or with human prompting, (6) infringe on one or more democratic norms (e.g., elections, freedom of speech), (7) cause or reinforces disparate impacts on protected classes of individuals, (8) threaten national security, and/or (9) substantially prevent achieving climate and sustainability goals, among others. An AI technology or specific system could be added to the list by regulators if, according to the latest scientific knowledge, adverse impacts on public interests or personal rights are expected. Incumbent upon reviewing agencies is promulgating rules to ensure its listing process is fair, accurate, transparent, and subject to judicial review. Regulators could provide a means for potential applicants to obtain an advisory opinion from regulators if they are uncertain whether their AI systems or technologies are one of the listed AI systems or otherwise fall within one of the above high-risk categories (the FDA currently offers this sort of pre-review under a program for medical devices).

Aside from the listed AI systems, covered persons meeting certain threshold criteria may, as noted above, also need to be regulated, an approach used in some privacy laws.[12] That is, as a proxy for potential risk, regulators may define covered persons by the number of users who have access to or are impacted by the covered person's AI system, or, if it is a company, by its size (measured by revenue generated by its AI technology).[13] Persons not meeting either the listing or other criteria may nevertheless need to be reviewed by an agency, perhaps because their architecture or purpose is new, or they collect, process, and monetize a significant new amount of user or other kinds of data used to train models, or because of some other significant change to their AI system (e.g., a new model architecture and dataset that affects system accuracy).

Permit application element

> Possible legislative language (definition): AI SYSTEM RISK IMPACT ASSESSMENT—means a study evaluating an AI system and its development process, including its conception, design, and training/testing datasets and methods, for impacts on accuracy, bias, discrimination, fairness, privacy, repeatability, and

> security, which includes, at a minimum—(A) a detailed description of the system, its design, development, purpose, and its actual and foreseeable uses directly or indirectly involving the public; (B) an assessment of the relative benefits and costs of the system in light of its purpose, taking into account relevant factors established by regulation; (C) an assessment of the risks posed by the uses of the system to the public; (D) the measures employed to prevent, minimize, and mitigate the risks described in subparagraph (C), including technological and physical safeguards used by the covered person who has a business or other interest in the AI system.[14]

A written permit application is the vehicle by which a covered person submits information to a reviewing agency in their effort to secure a permit. The application should contain all of the information a reviewing agency requests to fairly evaluate the applicant's AI system. This could include: a demonstration of compliance with applicable standards and criteria associated with its proposed uses, an identification of the system's benefits and whether they outweigh its risks, information about the system and how it works, information about the applicant's methods used in developing the system and the controls used to maintain its quality and continuous compliance with standards and criteria. The applicant should tell the system's "whole story," including answering questions like, what happened during training and testing, what are the components of the system, and what different ways may it be operated in the wild.[15]

A risk assessment is a common technique for prospectively demonstrating the risks of a particular action and whether social benefits of the action outweigh those risks.[16] As suggested in the statutory language above, covered persons could submit to agency reviewers a risk impact assessment modeled on enforceable representations about how an AI system will foreseeably be used in the wild. In the context of AI systems, social impacts could be measured against relevant public interests and individual personal rights that, if infringed, will have consequential adverse effects on a person's life.[17] The publication of appropriate technical and non-technical standards is crucial for both risk assessors and risk reviewers (agencies), a fact underscoring the important current work being done in this area by the U.S. National Institute of Standards and Technology (NIST) and others.[18]

For purposes of identifying who should apply for an AI system permit, lawmakers could assign that task to the covered person who actually "deploys" (directly sells) a system in the stream of commerce, in part because they are responsible for the decision to launch the system and can take the last measure of its benefits and whether it poses a risk of causing harm. With reference to Figure 15.2, developers are those who create AI systems using one or more AI and non-AI technologies, and deploy/sell them directly on a platform they themselves provide or on one provided by a third party accessible to users (e.g., a cloud-based platform). An example is a developer who publishes an app to a digital storefront

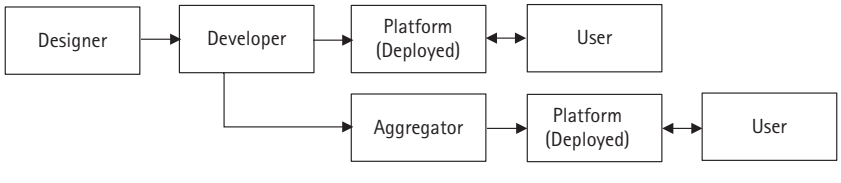

FIGURE 15.2 Two categories of responsible persons in an AI technology ecosystem

making it available for download and use on a smartphone (under current U.S. law, the digital storefront is generally not liable for the developer's app). A developer could also be a company that implements an internal AI system across its own internal network.

In other situations, a developer supplies its AI technology to a third party who combines it with others to create a system of systems, which is then deployed (directly sold) in a way that users access or interact with it. An autonomous vehicle is an example in which an AI technology may have a unique role that an aggregator combines with other systems. In that situation, the system aggregator, not the specific AI developer, may be the covered person who is responsible for seeking a permit.[19]

Application review element

Due to the nature of AI systems, it is expected that permit applications may need to be bifurcated into a "black box" portion, presumably reviewed by an institution with suitable machine learning technical expertise (such as NIST,[20] or a private entity certified by NIST or other suitable agency to conduct such reviews), and a risk and impact portion, which could be reviewed by a federal agency having jurisdiction over the general subject matter of the AI system described in a permit application (Figure 15.3).[21]

The technology review could include testing of applicant's datasets and models for things like accuracy, repeatability, generalization, apparent bias, data security, model leakage, and transparency, among other criteria. It is possible that this could be conducted using an application programming interface (API) provided by the applicant allowing reviewers direct access to applicant's models.

The non-technical review will require a reviewing agency to decide whether to permit an AI system, even if it poses some inherent risk to the public. Important for this evaluation is, as suggested, appropriate standards so an agency may properly assess whether, for example, an algorithmic decision system may be permitted to operate if, despite an applicant's demonstrating adequate risk mitigation measures, its technology is expected to make up to one wrong decision and 99 appropriate decisions. Eliminating risk altogether, by employing a no-risk standard, is presumably untenable.

Finally, it is suggested that some form of public participation (beyond citizen suits, discussed later) be an integral element of the AI permit program at least to provide transparency and thus improve trust in AI. This could include providing public review and an opposition comment

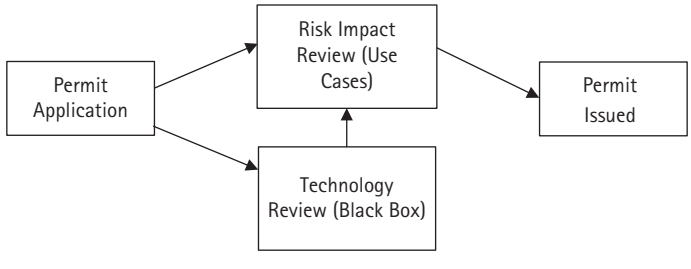

FIGURE 15.3 General permit application review process that bifurcates an application into two separate agency reviews.

period in which the public may provide comments to the reviewing agency before it makes its final determination and approves a permit application. The agency should be required to address all comments in a public forum. Concerns related to protecting trade secrets and confidential information would need to be addressed as part of any public review process.

Permit issuance element

A permit is a form of license agreement between a reviewing agency (and the public it serves) and a permit holder and take the form of permission to operate in exchange for continuous compliance with permit terms and conditions. As a legal instrument, an issued permit can reflect applicable normative rights and values (Brownsword et al., 2017). The permit consists of all applicable requirements in a single document that is accessible to regulators and to the public to ensure compliance. Its terms and conditions should be reasonable and unambiguous, expressing the metes and bounds of what an AI system is permitted to do and not do, and actions those responsible for the system must take under various circumstances.

Thus, it is expected that no permit should issue unless it includes conditions that will at least ensure compliance with all applicable legal requirements and mitigate risks. Permits should contain the plans and the schedules for covered persons to maintain compliance with applicable standards, plus limitations on the use of the system, monitoring requirements, and reporting provisions, including those covering the collection, analysis, use, disclosure, and sale of customer data.[22] Permits could include annual certification requirements, whereby a responsible party submits each year a report that certifies compliance with its permit. Table 15.1 identifies possible permit terms and conditions for AI systems.

Enforcement element

This section posits a public–private law enforcement approach is needed for an AI permit program, and briefly describes the elements thereof: a combination of agency-led law enforcement (including investigations), private rights of action, and citizen suits. As is the case in other regulatory permit systems, those who violate permits should face civil liability, as well as possible criminal penalties if their behavior is found to be negligent, intentional, or reckless.[23] It follows then that the primary purpose of enforcement under the AI permit program is to encourage compliance through the ever-present consequence of paying monetary damages or being subject to other significant remedies following violations. One way to achieve this is by providing a means for those who are harmed by permitted AI system to obtain monetary compensation, either through administrative fines collected by the reviewing agencies from violators (public law enforcement) or via private civil lawsuits brought against offending companies (private law enforcement) (Polinsky & Shavell, 2007). But enforcement elements could also include means for enforcing other relevant provisions of the AI permit program statute. Regardless of the specific approach used, applicable legal elements need to be sufficiently robust to encourage continuous compliance with the law and permit terms and conditions, dissuade risky behavior by those who operate permitted AI systems, and encourage development of AI systems that have a positive benefit on society. The contours of enforcement are described here, but due to limited space, the details are omitted.

Table 15.1 Exemplary terms and conditions reviewing agencies could include in AI permits, as applicable to the relevant technology underpinning an AI system

Category	"The company or person to whom this permit has been granted must at all times . . ."
Monitoring	Maintain an approved and up-to-date incident response plan[62] covering the permitted AI system.
Notices	Provide conspicuous notice to a user prior to user interaction with the permitted system, at the time of, and as soon thereafter as is reasonably possible.
Recordkeeping	Maintain records and make regular reports to the reviewing agency demonstrating on-going compliance with the permit's terms and conditions.
Recordkeeping	Maintain a registry for individuals harmed by the permitted system to report circumstances to authorities.
Recordkeeping	Maintain records of discrete decisions made by the permitted system that affect a customer for a period of five years sufficient for a finder of fact to understand how the decisions were made.
Reporting	Notify the agency when an entity behind an AI system ceases doing business, and whether its permitted system will persist or remain active.
Reporting	Notify the agency when the purpose for which the AI system is permitted is changed or is used by another for a different purpose.
Reporting	Notify the agency when a material change is made to the permitted system (e.g., changes to its model, datasets, purposes, etc.), and obtain supplemental review and approval prior to deploying the changed system.[63]
Reporting	Report to the agency and any other authority stated in the permit any known incidents of harm caused by the permitted system as soon as they are known to the permittee.
Representations	Report to the reviewing agency when the permittee knows that it has not or cannot meet the representations it made in its permit application, which are incorporated into and forms a part of the issued permit.
Technical	Continuously meet all applicable standards immediately upon launch and within a prescribed time period after new standards are published.
Technical	Maintain an accuracy as good as the best similar systems already approved for similar machine learning-based models and algorithmic decision-making techniques.

Agency law enforcement

> Possible legislative language: The Administrator shall enforce permits, permit fee requirements, and the requirement to obtain a permit, including authority to recover civil penalties in a maximum amount of not less than [$10,000] per day for each violation, and to seek appropriate criminal penalties.

As suggested by this example statutory language, civil fines could be tied to degrees of noncompliance, similar to other permit programs. That is, a significant fine could be imposed

for significant violations, smaller amounts could be imposed (often, per day) for minor violations (until the noncompliance is corrected). In cases of knowing and negligent endangerment of others due to the actions by the permit holder in its operation of an AI system, an agency could treat such actions as criminal.[24] In the criminal context, penalties could be tied to the degree of intent.

Reviewing agencies should be given the power to conduct investigations (which requires subpoena powers).[25] In the case of AI systems, remedies for some violations could include orders to eliminate the source of the harm (e.g., an order to take down [un-deploy] an AI system) and/or delete/destroy datasets and models. This is precisely the approach the U.S. Federal Trade Commission (FTC) chose as part of its enforcement of deceptive trade regulations against a facial recognition company in early 2021.[26]

Citizen suits

> Possible legislative language: Any person may commence a civil action on their own behalf: (1) against an agency administrator where there is alleged a failure to perform any act or duty that is not discretionary, or (2) against any person who proposes to deploy any AI system without a required permit.[27]

Goodwin (2020) points out that agency enforcement personnel cannot observe all regulatory violations within their respective jurisdictions, hence the value of citizen suits. Citizens (as well as other forms of public participation) can serve as an agency's "eyes" and "ears," and citizen suits permit citizens to vindicate harms to the general public. In that way, lawmakers can avoid the prospect of court dockets being overwhelmed with actions by "every subject in the kingdom" who decides to sue "the offender with separate actions."[28] One of the barriers to bringing citizen suits on behalf of others is standing, which, in the area of AI technologies, is unsettled (Oberly, 2021). Standing is a particularly contentious issue in the context of AI systems because potential harm may not be manifest until long after an incident occurs.[29]

The suggested statutory language above provides that citizen suits may also be used to hold administrators accountable when their actions are viewed as contrary to law. That is, citizen suits allow for public suits against administrators for (non-discretionary) decisions that are ill-conceived, fall short of expectations, or appear counter to the law.[30]

Private right of action (private remedy)

> Possible legislative language: Any person aggrieved by a violation of this Act may commence a civil action on their own behalf in a federal district court against any offending party who is alleged to have violated or to be in violation of this Act, an implementing regulation, or a standard, condition, limitation, or representation expressed in a permit issued under this Act.[31] Any such violation constitutes an injury-in-fact and a harm to any affected individual.[32] An offending party may not, as a basis for seeking to dismiss such a civil action under this paragraph, rely on an arbitration provision unless the party bringing the action had been given prior notice of the provision and affirmatively indicated their consent using an opt-in mechanism that recorded their opt-in choice.

In 2008, lawmakers in Illinois chose to include a private right of action in their Biometric Information Privacy Act (BIPA), which since then has been cited by plaintiffs

in hundreds of privacy lawsuits against companies that make or use biometric data collection systems in the state, many of which are powered by AI technologies (e.g., machine learning for facial recognition). Though controversial, the private remedies provision has been credited with increased compliance with the law.[33] Federal lawmakers are considering a federal biometric privacy law having a similar private right of action provision.[34]

As BIPA court cases demonstrate, standing in AI technology cases remains in flux, in part because of questions about whether a plaintiff must demonstrate actual harm when a defendant violates a statutory provision. Consider, for example, an AI-operating permit condition requiring a permit holder to post notice that a person's biometric identifier or biometric information is being collected or stored. If a plaintiff alleges no such notice was provided to them, but their showing of actual personal harm is tenuous, a court may dismiss plaintiff's complaint on the basis of lack of standing despite the defendant's unmistakable law violation. For lawmakers, this may be addressed by expressly stating that violations constitute an injury-in-fact and a harm to plaintiff, as noted in the proposed statutory language above. The practical effect of this is to increase the odds the plaintiff's case will survive an early challenge by defendants.

Lawmakers may also choose to provide for awards of reasonable attorney's fees and court costs to prevailing parties as a means to lower the barrier of access to the courts for those with meritorious cases but who otherwise may not sue due to a lack of resources to bring a cause of action themselves. Lawmakers may also wish to address the use of arbitration agreements buried in terms of use agreements (that most users never read) to defeat civil suits.[35]

Judicial review

> Possible legislative language: Any action respecting a violation of a permit standard, limitation, or condition, or an order respecting such standard, limitation, or condition, may be brought in the judicial district where the harm occurred.

In terms of venue, lawmakers could allow actions in any federal district court where the act that gave rise to the permit violation occurred or where the harm arose, and without regard to the amount in controversy or the citizenship of the parties, thus potentially increasing the ability of plaintiffs suing where they reside. Because highly technical issues are expected to arise in citizen suits and private litigation involving AI technologies, a single forum for appeals could provide greater uniformity in interpretation of laws and reviewing non-discretionary agency decisions (for example, published opinions of a single, national appeals court would apply nationally rather than being limited to a particular regional circuit). An existing forum for appeals in the U.S. is the Court of Appeals for the Federal Circuit (CAFC).[36] This court was established in 1982 following recognition of then-existing problems associated with the lack of uniform rulings in specialized areas of the law. The CAFC, for example, handles all patent appeals due to the technical and specialized area of law applicable to patents. The CAFC's national jurisdiction could be expanded to include AI technologies by amending existing law.[37]

The Politics of Permits: Government Intrusion, Federalism, and the Status Quo

Although the approach taken above is intended to be primarily a pragmatic one, it is instructive to briefly consider some of the political dimensions of a significant new regulatory permit program for AI technologies. Two issues in particular—government intrusion and federalism—are expected to be mentioned if debate ensues. Both would be measured against the status quo, which in the U.S. may be defined by a combination of existing statutory laws—civil rights, labor and employment, consumer protection, and privacy; agency regulations and relevant common law; and private civil tort litigation.[38] Critics of government intrusion may point out the existence of these alternatives as being far less burdensome on the AI industry than a new permit program, and they would be correct too that the alternatives arguably provide some of the benefits of an *ex ante* permit program.

Consider, for example, the Federal Trade Commission Act, which authorizes the FTC to regulate unfair competition and deceptive trade practices, including those that AI companies might engage in. For some injured parties, the FTC's existing regulatory program may serve them well, in part because the agency has law enforcement powers (including subpoena authority) and can penalize those who caused their injuries (in some cases, paying the injured directly from the money receive from enforcement actions).[39] But, as a form of "social insurance" (Holmes, 1881, p. 96), these enforcement-led approaches necessarily rely on knowledgeable administrators who can set appropriate expectations via published policies or rulemaking to curb risky behavior. Being unaware of the nuances of AI technologies and rapid changes thereto may lead regulators to promulgate the wrong (or no) expectations. Coupled with spotty or ineffective enforcement, some companies may not be deterred, the potential result being incomplete protection of the public's interests and private individual rights.[40] Although specific permits also contain normalized expectations (embodied in permit terms and conditions), they can be more tailored to specific companies and their AI systems, avoiding the one-size-fits-all approach of alternative non-permit approaches.

Private tort litigation, as Morgan (2017) points out, can complement agency enforcement by its deterrent effect on negligent behavior.[41] In tort litigation, an individual seeks remedies in civil law against a person who commits a tortious act that result in loss or harm. But, as seen, permitting agencies with subject matter experts may be better equipped to deal with a wide-range of technical subject matter and perform a "broad and sophisticated brokering of social costs and benefits"[42] that judges in tort litigation, who are confined to the specific facts of a case, may not reach.

If existing laws and regulations already provide a net benefit to society, critics of an AI permit program would ask whether its added social benefit justifies additional government intrusion. As Justice Holmes famously wrote in 1881, "State interference is an evil, where it cannot be shown to be a good."[43] In the context of an AI permit program, that postulate can properly be answered only after a full economic and societal impact assessment (which the Congressional Budget Office (CBO) and designated AI permit reviewing agencies would have to complete prior to an AI permit law being enacted).[44] For these purposes, the apparent additional good produced by an *ex ante* AI permit program may be expressed in

terms of the avoidance of harm that otherwise unpermitted AI technologies might cause. The other side of Holmes' coin would consider the economic impact of a new regulatory permit program on the regulated AI community, which could be significant.

The other debate issue—federalism—assumes government interference is necessary, then asks the question, should federal lawmakers give state or local governments the power to decide whether and how they will regulate AI technologies operating within their respective jurisdictions? Looking at it from an historical perspective, Reitze (1999) points out that EPA's federal permit program for stationary air pollution sources arose due to failures of state and local governments to effectively deal with air pollution at local and regional levels. Before the EPA's federal air permit program, an industrial source's pollution control obligations were scattered across numerous state plans or various federal regulations.[45] Approximately 35 states or localities had attempted operating permit programs, but few were as comprehensive as the EPA's national program that replaced them.[46] Indeed, Congress's intent in creating a national permit program for stationary sources included better enforcement of the requirements of the law by applying them more clearly to individual sources and allowing better tracking of compliance, as well as expedited implementation of new pollution controls after they were developed.[47] Even so, lawmakers in the end crafted the program that gives authority to states, local agencies, and tribes to issue permits, with the EPA providing oversight and final approval only if they fail.[48] That is partly because, at the time of its enactment, the EPA reportedly did not have the people or money to run a national permit program.[49]

Arguably, today's human talent deficit in AI[50] suggests giving states more say in permitting AI technologies in their jurisdictions. But others could just as persuasively argue that a far better use of scarce resources is to combine available talent within a few federal agencies[51] rather than having potentially dozens or hundreds of local agencies competing against one another (and with the technology companies) for data scientists, machine learning software engineers, and other AI subject matter experts.

Those in favor of a federal program can also point to its avoidance of the jurisdictional race to the bottom problem, characterized by companies choosing to operate in jurisdictions where the regulatory burden on them is lower, or treating customers differently depending on where they live. This is not a theoretical problem for AI. In a 2021 FTC enforcement action, the agency found a facial recognition company's deceptive data use actions targeted Americans who lived in states with no specific state law protections.[52] According to the FTC's complaint, the company, for a period of time, used one opt-in approach for customers located in Illinois, Texas, and Washington (states with biometric information data collection and data privacy laws in place) and another for individuals residing elsewhere.

A federal program is also seen as better suited to addressing national security concerns. All things being equal, local governments may not have access to the intelligence that central governments do, and any response they muster may have only local effect. Additionally, local administrators may not have the infrastructure and resources in place to perform the kinds of detailed intelligence reviews of technologies that the federal government can.[53]

Notwithstanding, there may be legitimate reasons why local, state, or regional administrators may want to impose greater restrictions on certain AI systems that impact citizens in their respective jurisdictions. Thus, lawmakers may choose not to categorically preempt local AI permit programs, or other local laws that touch on AI technologies, as long as they

impose on local authorities certain minimum standards and requirements for their local permit programs.[54]

Europe's Proposed AI Regulations, a U.S. Program for Medical Devices, and the "General" Permission Regulatory Approach

As the work for this chapter was being completed, the European Commission (EC) published its draft framework for regulating high-risk AI systems and practices in the European Union.[55] That proposal and the permit program suggested here broadly fall under the rubric of permission-type governance schemes, though the EC's proposal is fairly characterized as a "general approval" process given its lack of specific/individual permits. By comparison, the present permit program follows a process Biber and Ruhl (2015) describe as, an "agency engaging in extensive fact gathering and deliberation particular to the individual circumstances of an applicant's proposed action, after which the agency issues a detailed permit tailored just to that applicant's situation."[56] Table 15.2[57] identifies some of the similarities and differences between the EC's proposal and the present one. Where they diverge the most is in their enforcement mechanisms, as the EC's proposal contains no private right of action, citizen suit, and single appellate jurisdiction provisions.

While much too early to predict whether general approvals or specific permits will be more effective at reducing AI risks, at least one general approach used in the U.S. for AI-powered medical devices has its share of critics. In particular, the FDA's medical device pre-market approval (PMA) regulatory program, which evaluates technologies, including software, against safety and effectiveness standards,[58] has been criticized as lacking transparency. As a general approval scheme, PMA applies to devices according to risk: higher risk systems undergo full PMA reviews by the FDA, consisting of a detailed technical submission review, while lower risk devices undergo de novo reviews and approval if similar technologies have previously been approved and "general controls alone, or general and special controls, provide reasonable assurance of safety and effectiveness for the intended use."

Benjamens et al. (2020) point out that the agency's current de novo reviews of medical devices generally lacks transparency (from the public's perspective), and the FDA's announcements of approved new systems also do not provide clear descriptions of the systems, nor do they permit third parties the ability to assess the implementation of the devices. Ross (2021) found that few of the AI-powered products cleared by the FDA include public summaries filed by manufacturers that provided transparency about the racial demographics of data the manufacturers used when assessing algorithmic performance.

With the rise in AI medical technologies, or so-called software-as-a-medical device (SaMD; defined as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device"), the FDA's de novo review process may be modified to include a pre-certification (Pre-Cert) procedure for SaMD,[59] which may include evaluating the companies proposing to market a SaMD device using so-called Excellence Appraisals (a review of whether a company meets "excellence

Table 15.2 Comparing the legal elements identified in this work with Europe's proposed regulations

Legal Element	This Chapter	European Commission Proposal
Permission structure	Individual/specific permission	General permission
Authority	Federal law; federal regulatory administration	Individual EU members and their administrative ("notified") bodies
Applicability	i. Listed high-risk systems ii. Higher-risk companies and natural persons iii. "Other" (lower risk)	i. Listed high-risk systems and practices ii. Moderate-risk systems and practices
Exclusions?	None	AI systems developed and used exclusively for military purposes
Prohibited classification?	None (but some technologies may not be permitted to operate)	Yes: "prohibited practices"
Requirements for "other"	E.g., registration, certification	E.g., voluntary adoption of practices ("codes of conduct")
Regulatory submission	Formal permit application, including technical description and prospective AI risk impact assessment	Technical Documentation submission, including risk analysis and *ex ante* conformity analysis
Submission review	Administrative regulatory agency (bifurcated analysis)	Administrative notified bodies in Union member where AI system deployed
Grant conditions	Applicant-specific recordkeeping, reporting, and other requirements	Entity-specific approval upon showing of legal conformance, risk management measures, and other requirements
Safe harbor?	Yes (limited, including real-time testing)	Yes (for testing purposes; regulatory sandbox)
Public participation?	Yes (e.g., oppositions to a proposed permit approval)	Not specifically provided for
Regulatory violation considered injury-in-fact?	Yes	Not specifically provided for
Administrative enforcement	Penalties (corrective action, fines, take down)	Penalties (corrective action, fines, take down)
Private right of action?	Yes	Not specifically provided for
Citizen suits?	Yes	Not specifically provided for

principles" of patient safety, product quality, clinical responsibility, cybersecurity responsibility, and exhibits a proactive culture and ability to deliver safe and effective software products). This new approval process would include obtaining commitments from manufacturers to provide transparency and "real-world performance monitoring" of cleared devices.[60] These organization-based assessments may include rating the quality of a

manufacturer's software design, its testing of its software/device, how it conducts real-world monitoring (which presumably relates back to assessing the core performance standards of safety and efficacy), and other capabilities.[61] Whether these proposed changes will satisfy critics is an open question. The American Medical Association for its part suggested that, for purposes of improving transparency and trust, SaMD systems contain an "Augmented Intelligence Facts" label, similar in concept to a drug facts label, to inform the public of potential AI system risks.

Conclusion

AI technologies, systems, and practices are today navigating an uncertain legal landscape, one characterized more by its lack of targeted laws and regulations as much as anything. This creates risks for companies and individuals operating in global AI markets and leaves portions of society vulnerable to potential harms caused by AI. To mitigate potential harms, lawmakers can choose among several proposed governance frameworks for AI, including a national regulatory permit program of the type suggested here, which is arguably potentially more protective of public interests and individual rights, but also more burdensome on administrators and the regulated community than some of the alternatives.

Notes

1. Morgan (2017, p. 526).
2. Bell & Ibbetson (2012, p. 164).
3. Moses (2017, p. 588), citing Collingridge (1980), who wrote, in the context of technology, "When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming."
4. See also Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence (2020). No U.S. federal legislation introduced since 2020 was identified as proposing permitting as a means for regulating AI.
5. H.R. 153 (2019).
6. See Hudson (2021; chapter 3.1 of this title) for a discussion of the interests and rights at stake in AI. Infringement of some of these rights could cause irreversible harm (Moses, 2017, p. 579). See also *Spokeo, Inc. v. Robins*, 136 S. Ct. 1540, 1551 (2016) (explaining that personal rights belong to the individual, considered as an individual, including rights of personal security, property, contract, and privacy).
7. Toews (2020) suggested adapting existing national permit programs to the governance of AI. Biber and Ruhl (2015) provide the federal government's perspective and framework for agencies to think about when creating regulatory permit programs.
8. Adapted from earlier work, "AI technologies" are defined as the software-implemented algorithms created using machine learning techniques and other approaches that analyze large datasets to create digital models of the real world. "AI systems" are the software and related hardware deployed by AI developers to operationalize the digital models in a private or public enterprise (Higgins & Dodd 2021, p. 1). "AI" is used as shorthand for both definitions, depending on the context.

9. In fairness, the law could exempt AI systems that exist prior to the law's effective date, but at the same time could also make the law apply to those systems if their owners make substantial "changes" (which will need to be defined in the law).
10. *See* Food and Drug Administration Modernization Act of 1997, as amended, by which Congress created a tiered risk classification scheme (Class I, II, and III) for medical devices.
11. *See* Clean Air Act Amendments of 1990, Sections 108, 109 and 42 U.S.C. §7412.
12. California Consumer Privacy Act (CCPA); Virginia Consumer Data Protection Act (VCDPA).
13. *See* Senate Bill 2637 (October 17, 2019).
14. *See* Senate Bill 2637 (October 17, 2019).
15. This kind of information is analogous to the information the FDA seeks from applicants wishing to market a new drug in the U.S.
16. Clean Air Act Amendments of 1990 Conference report filed in House (October 26, 1990), explaining Title V (permits): "Authorizes the issuance of a permit to construct or operate a new source if demonstrated that the benefits of such source significantly outweigh environmental and social costs." Also, see National Security Commission on Artificial Intelligence, Final Report (March 1, 2021), Chapter 8, which suggests the government prepare AI Risk Assessment Reports and AI Impact Assessments to assess the privacy, civil liberties and civil rights implications for each new qualifying AI system or significant system refresh, as part of improving transparency about AI use.
17. When reviewing drugs/medical devices, the FDA looks at whether a device's use is of substantial importance in preventing impairment of human health, or if the device presents a potential unreasonable risk of illness or injury. See FFDCA, 21 U.S.C. §860.3(c)(1)–(3).
18. It has been suggested that risk assessments be conducted by independent assessors, paid for by the applicants seeking permits. Under the program proposed here, regulatory authorities (or their commercial vendors) would have the burden to independently evaluate the sufficiency of an applicant's risk assessment submission.
19. Complex legal questions about liability, joint and severable liability, agency, contractual shifting of liability, disclaimers, and related concepts are relevant and should be given consideration.
20. NIST is a logical choice for black box technology reviews as it has for many years engaged in the testing of machine learning systems, including those of government vendors as part of its Face Recognition Vendor Test (FRVT) series (2000–present) (see https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt).
21. By way of example, the FDA could review permit applications for AI systems whose purpose is diagnosing or treating disease and the Consumer Financial Protection Bureau could review permit applications directed to systems related to banking, securities exchanges, and other financial matters. A single agency to review all AI technologies and systems is also possible, with subject matter experts gathered under one roof (such as the U.S. Patent and Trademark Office, with its corps of patent Examiners who are assigned to various "art units" depending on their individual areas of technical expertise).
22. *See, e.g.,* Title 40, Code of Federal Regulations (CFR), §71.6 (describing content of permits issued under Title V of the Clean Air Act Amendments of 1990); Title 21, CFR, §814.82 (describing content of approvals issued under the Federal Food, Drug, and Cosmetic Act).

23. As part of a company's overall legal risk management, an AI incident response plan, like a data exfiltration response plan that many companies already maintain, identifies actions a company will take (including notifying appropriate administration officials) once it becomes aware that its actions or its AI systems have violated a permit condition, harmed a user or the public, or caused some other adverse impact.
24. What constitutes a "change" is a significant area of concern for both regulators and the regulated, given that some AI systems are updated frequently, for example by altering the network architecture to address state of the art developments and/or by retraining models with updated "live" data (which may be a necessity to meet accuracy and other technical standards set forth in permits). An approach the FDA uses to assess change is to define alterations that "significantly" affect health and safety, and those that are "major changes or modifications" in the intended use. *See* FDA "Guidance for Industry and Staff: Deciding When to Submit a 510(k) for a Change to an Existing Device" (2017). https://www.fda.gov/media/122535/download.
25. By comparison, the EPA may enforce a Title V permit either administratively or in federal court. 42 U.S.C. § 7413(a), (b).
26. See Title VII of the CAAA of 1990.
27. See FTC Act Sec. 3, 15 U.S.C. § 43.
28. *In re Everalbum, Inc.*, No. 1923172, Agreement and Consent Order (Jan. 11, 2021). https://www.ftc.gov/system/files/documents/cases/everalbum_order.pdf.
29. Examples: 42 U.S.C § 7604 (air pollution); 33 U.S.C § 1365 (water).
30. *Spokeo, Inc. v. Robins*, 136 S. Ct. 1540, 1551 (2016).
31. *See Lopez, et al. v. Apple, Inc.*, slip op. 19-cv-04577 (N.D. Cal. Feb., 10, 2021) (granting motion to dismiss on grounds that plaintiff's alleged harm from unintended *Siri* voice recordings sent to third party was too speculative). Followers of standing in AI technology cases are closely watching courts in Illinois, which are considering standing under the state's Biometric Information Privacy Act (BIPA), 740 ILCS 14/20 (Oct. 3, 2008).
32. Based in part on 5 U.S.C. § 702, Administrative Procedures Act (APA) ("A person who can demonstrate they suffered a legal wrong, including harm, because of an agency action, or adversely affected or aggrieved by agency action within the meaning of a relevant statute, is entitled to judicial review thereof.").
33. Based in part on BIPA, 740 ILCS 14/1 *et seq.*
34. Based in part on S.4400 (August 3, 2020).
35. *See In re Everalbum, Inc.*, Agreement and Consent Order (finding that Everalbum, for a period of time, used one opt-in approach for customers located in Europe, Illinois, Texas, and Washington—states with biometric information data collection and data privacy laws in place—and another for customers located elsewhere).
36. S.4400 (August 3, 2020).
37. *See Wiilcosky v. Amazon, Inc.*, 2021 WL 410705 (N.D. Ill. Feb. 5, 2021) (dismissing biometric privacy lawsuit on the grounds that by creating Amazon accounts and making purchases with their accounts, plaintiffs consented to Amazon's arbitration agreement contained in the company's Conditions of Use agreement).
38. Public Law. 97-164, Federal Courts Improvement Act, 96 Stat. 25 (1982).
39. See 28 U.S.C § 1295, Jurisdiction of the United States Court of Appeals for the Federal Circuit, which could be modified by adding new paragraph (15): "The United States Court of Appeals for the Federal Circuit shall have exclusive jurisdiction— * * * (15) of an

appeal from a final decision of a district court of the United States . . . in any civil action arising under, or in any civil action in which a party has asserted a compulsory counterclaim arising under, any Act of Congress in which harm caused or contributed by an AI technology or AI system is alleged."

40. Similar legislation and directives exist in the European Union. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
41. So-called "compensation funds" are a technique used by federal agencies and departments, including the FTC.
42. Examples include the levying of large agency fines failing to alter behavior. After the FTC imposed a $5 billion fine on Facebook following the Cambridge Analytics incident (in which Facebook permitted private user data to be accessed by a third party without user notice), Facebook has since upped its mining of user data to train its AI systems. *See* "Learning from videos to understand the world," *Facebook AI* (March 12, 2021). https://ai.facebook.com/blog/learning-from-videos-to-understand-the-world.
43. Morgan (2017, p. 536, 540). Scherer (2018), Moses (2017), and Reitze (1999) examine other benefits and problems with tort litigation involving existing and new technologies.
44. Morgan (2017, p. 537).
45. Holmes (1881, p. 96).
46. As per the Congressional Budget and Impoundment Control Act of 1974, and Executive Order 12866 (as amended).
47. Congressional Research Service, "Clean Air Permitting: Implementation and Issues" (Sep. 1, 2016). https://crsreports.congress.gov/product/pdf/RL/RL33632.
48. Id. at 1
49. Id.
50. Under the federal Clean Air Act Amendments of 1990, states are responsible for issuing Title V permits. 42 U.S.C. § 7661a (b), (d).
51. Reitze Jr. (1999).
52. *See* National Security Commission on Artificial Intelligence, Final Report (March 1, 2021), Chapter 6.
53. Examples of this are numerous, but one, in particular, stands out: the U.S. Patent Office, with its corps of patent examiners operating under a single agency roof.
54. *In re Everalbum*, No. 1923172, Statement of Commissioner Rohit Chopra (January 8, 2021), at p. 2.
55. By way of example, the U.S. Patent Office can screen all new patent applications for possible national security concerns. 35 U.S.C. §17.
56. State-specific data privacy laws, for example, should not be set aside.
57. European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (April 21, 2021).
58. Biber and Ruhl (2015, p. 2).
59. Table 15.2 is informational only; no legal determinations should be made based solely on the information provided.
60. Aspects of the program may be found in provisions of several laws, including the Federal Food, Drug, and Cosmetic Act (FDC&A), FDA Modernization Act (FDAMA), FDA Safety Innovation Act, 21st Century Cures Act. See also, Digital Health Software Precertification Program (Pre-Cert). https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program.

61. FDA, "Artificial Intelligence and Machine Learning in Software as a Medical Device." https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
62. FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback" (April 2, 2019). https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program.
63. FDA, "Developing Software Precertification Program: A Working Model" (June 2018). https://www.fda.gov/media/113802/download.

References

Biber, E., & Ruhl, J. (2014). The permit power revisited: The theory and practice of regulatory permits in the administrative state. *Duke Law J. 64*(2), 133. http://scholarship.law.duke.edu/dlj/vol64/iss2/1.

Biber, E., & Ruhl, J. (2015). Designing regulatory permits: Report and case studies. Administrative Conference of the United States. https://www.acus.gov/sites/default/files/documents/Licensing%20and%20Permitting%20Draft%20Report.pdf.

Bell, J., & Ibbetson, D. (2012). *European legal development: The case of tort*. Cambridge University Press.

Benjamens, S., Dhunnoo, P., Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digital Medicine 3*

Black, H. C. (1990). *Black's Law Dictionary*. West Pub. Co.

Brownsword, R., Scotford, E., & Yeung, K. (2017). Law, Regulation, and Technology: The Field, Frame, and Focal Question. In Brownsword, R., Scotford, E., & Yeung, K. (Eds), *Oxford handbook of law, regulation, and technology* (pp. 15–20). Oxford University Press.

Collingridge, D. (1980). *The social control of technology*. Frances Pinter.

DeMuth, C. (2012). The Regulatory State. *National Affairs 50* (summer), 75.

Fischer S.-C., Leung J., Anderljung M., O'Keefe C., Torges S., Khan S., Garfinkel B., & Dafoe A. (2021) "AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment". (2021) #2021-10, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.

Goodwin, J. (2020, September). Citizen suits are good for the regulatory system, and we need more of them. *The Center for Progressive Reform*. https://www.progressivereform.org/cpr-blog/citizen-suits-are-good-regulatory-system-and-we-need-more-them/.

Higgins, B., & Dodd, A. (2021). Artificial intelligence and trust: Improving transparency and explainability policies to reverse data hyper-localization trends. *J. of Science and Law* 8(1), 1–11.

Holmes, O. W. (1881). *The common law*. Little Brown.

Hudson, V. (2021). Standing up a regulatory ecosystem for governing AI decision-making: Principles and components. In *Oxford handbook of artificial intelligence*. Oxford University Press.

Morgan, J. (2017). Torts and technology. In Brownsword, R., Scotford, E., & Yeung, K. (Eds), *Oxford handbook of law, regulation, and technology* (Chapter 22, pp. 578–581, 586, 588). Oxford University Press.

Moses, L. (2017). Regulating in the face of sociotechnical change. In Brownsword, R., Scotford, E., & Yeung, K. (Eds), *Oxford handbook of law, regulation, and technology* (Chapter 24, pp. 578–581, 586, 588). Oxford University Press.

Oberly, D. (2021, April). Personal jurisdiction challenges in BIPA class actions. *Bloomberg Law*. https://www.blankrome.com/sites/default/files/2021-04/bloombergperspectives_-_blankromeapril2021_-_oberly.pdf.

Polinsky, M., & Shavell, S. (2007). The theory of public enforcement of law. In A. Mitchell Polinsky and Steven Shavell (Eds.), *Handbook of law and economics*, *volume 1* (Chapter 6, pp. 406–407). Elsevier B.V.

Reitze, A. (1999). The legislative history of U.S. air pollution control. *Houston L.R. 36* , 679.

Ross, C. (2021, February). As the FDA clears a flood of AI tools, missing data raise troubling questions on safety and fairness. *Stat*. https://www.statnews.com/2021/02/03/fda-clearances-artificial-intelligence-data/.

Scherer, M. (2018). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard J. of Law and Tech. 29*, 354.

Toews, R. (2020, June 28). Here is how the United States should regulate artificial intelligence. *Forbes*. https://www.forbes.com/sites/robtoews/2020/06/28/here-is-how-the-united-states-should-regulate-artificial-intelligence/?sh=5bb789337821.

Legislative, Judicial, and Executive Materials

California AB-375, California Consumer Privacy Act of 2018 (as amended by the California Privacy Rights Act of 2020) (codified at Cal. Civil Code, Title 1.81.5, Part 4, Div. 3).

Congressional Research Service, "Clean Air Permitting: Implementation and Issues" (Sep. 1, 2016). https://crsreports.congress.gov/product/pdf/RL/RL33632.

Illinois Biometric Information Privacy Act (BIPA) (Oct. 3, 2008) (codified at 740 ILCS 14/1 *et seq*.).

In re Everalbum, Inc., Case No. 1923172, FTC Agreement and Consent Order (Jan. 11, 2021).

Kuznik v. Hooters of America, *LLC*, slip op., 2020 WL 5983879 (C.D. Ill. Oct. 8, 2020).

Miracle-Pond v. Shutterfly, Inc., slip. op., 2020 WL 2513099 (N.D. Ill. May 15, 2020).

Public Law 38 Stat. 717, Federal Trade Commission Act of 1914 (codified at 15 U.S.C. § 41 *et seq*.).

Public Law 52 Stat. 1040, Federal Food, Drug, and Cosmetics Act of 1938 (codified at 21 U.S.C. § 301 *et seq*.).

Public Law 82-256, Invention Secrecy Act of 1951 (codified at 35 U.S.C. § 181 *et seq*.).

Public Law 89-554, Administrative Procedures Act, Sec. 702 (Right of Review) (codified at 5 U.S.C. § 702).

Public Law 93-334, Congressional Budget and Impoundment Control Act of 1974 (codified at 2 U.S.C. § 601 *et seq*.).

Public Law 97-164, Federal Courts Improvement Act of 1982 (96 Stat. 25) (1982).

Public Law 101-549, Clean Air Act Amendments of 1990 (codified at 42 U.S.C. § 7401 *et seq*.).

Letter from J. Madara (AMA) to N. Sharpless (FDA), re "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), Discussion Paper and Request for Feedback" (2019 June 3).

Lopez v. Apple, *Inc*., slip op. 19-cv-04577 (N.D. Cal. Feb., 10, 2021).

National Security Commission on Artificial Intelligence, Final Report (2021 Mar. 1) https://www.nscai.gov/2021-final-report/.

Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1551 (2016).
U.S. FDA, Digital Health Software Precertification (Pre-Cert) Program. https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program (accessed April. 12, 2020).
U.S. FDA, "Artificial Intelligence and Machine Learning in Software as a Medical Device" (Jan. 2001). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
U.S. FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback" (2019 Apr. 2).
U.S. FDA, "Developing Software Precertification Program: A Working Model" (2018 Jun.). https://www.fda.gov/media/113802/download.
U.S. FDA "Guidance for Industry and Staff: Deciding When to Submit a 510(k) for a Change to an Existing Device" (2017). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-change-existing-device.
U.S. House of Representatives, H.R. 153, "Supporting the Development of Guidelines for Ethical Development of Artificial Intelligence" (2019).
U.S. House of Representatives, "Conference Report: Clean Air Act Amendments of 1990" (1990 Oct. 26).
U.S. Senate, S.B. 2637, "Mind Your Own Business Act of 2019" 116th Congress (2019 Oct. 17).
U.S. Senate, S.B. 4400, "National Biometric Information Privacy Act of 2020" 116th Congress, (2020 Aug. 3).
Virginia Consumer Data Protection Act (VCDPA) (2021) (codified at Va. Code Ann. § 52).
White House, Executive Order 12866, "Regulatory Planning and Review" (1993 Sep. 30) (as amended) (published at 58 Fed. Reg. 190 (Oct. 4, 1993)).
White House, Office of Management and Budget, "Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications". Russell T. Voigt. (2020, Nov. 17).
Wiilcosky v. Amazon, Inc., 2021 WL 410705 (N.D. Ill. Feb. 5, 2021).

CHAPTER 16

AI LOYALTY BY DESIGN

A Framework for Governance of AI

ANTHONY AGUIRRE, PETER B. REINER, HARRY SURDEN, AND GAIA DEMPSEY

Today, artificial intelligence (AI) systems are used in a variety of autonomous tasks, from searching and scheduling, to prediction and provision of advice. Accordingly, much has been written about fairness, accountability, and transparency in the context of AI use. But largely missing from this conversation is the concept of "AI loyalty." The issue of AI loyalty arises because of a bifurcation between those who develop AI systems and those who ultimately use them. AI systems are often created by corporations or other organizations, but the end-users (or beneficiaries) are often distinct individuals. The possibility of conflict of interest therefore arises: such AI systems may provide automated results that benefit the creators of such systems (or third parties) at the expense of end-users. For example, an AI-enabled medical advisory system may subtly recommend pharmaceutical drugs to doctors that are more profitable for a hospital system intermediary but that come at the expense of patient physical or financial well-being. In other cases, AI-enabled search or recommendation engines may appear to provide the best results for end-users while implicitly prioritizing the interests of AI creators. To illuminate, and ultimately reduce, such scenarios, in this chapter we develop the concept of "AI loyalty": this is the principle that AI systems should be designed, from the outset, to primarily and transparently benefit their end users, or at minimum clearly communicate conflict-of-interest tradeoffs, if they cannot be eliminated. The principle of "AI loyalty" by design becomes especially important when AI systems are deployed in contexts that today involve high degrees of trust, vulnerability, and ethical responsibility, such as between a doctor and patient, attorney and client, judiciary and population, and other professional and fiduciary settings.

Introduction

A techno-social ecosystem that includes artificial intelligence (AI[1]) fundamentally depends upon public trust. Accordingly, researchers have put forth persuasive arguments highlighting the importance of fairness, accountability, transparency and other trust-related values in

the use of AI (Lepri et al., 2018). Missing from this conversation is whether AI systems will act in the interests of their *end users*, or in the interest of others. Put another way, to whom will AI systems be loyal?

Consider the following. Imagine hiring a personal assistant via a recruitment firm to help organize your appointments, files, correspondence, etc. The assistant does excellent work, and the cost is reasonable. But after a month you read the fine print of the contract, discovering that the assistant is still technically employed by the recruitment firm and not you. Moreover, the contract requires the assistant to secretly send the firm summaries of all your private correspondence, finances, preferences, and private interactions so that the firm can use this information in its business. The recruitment firm is allowed to sell your personal information for profit or use it to persuade you to buy various products from them. Would this feel appropriate, or would it feel disloyal—as a betrayal even?

This is roughly the situation in which we find ourselves today with many AI tools and their attendant information systems. As we eye the increasing role of AI deployed by corporations with a view toward profits, a central question will be: will the design of AI systems be sensitive to the needs, wants, and values of *users*, or will they only *appear* to be loyal to users while actually maximizing gains for their corporate creator, or a powerful intermediary such as a bank, hospital system, or judicial system? In earlier work, we suggested that focusing upon conflicts of interest, and hence loyalty, brings clarity to the situation (Aguirre et al., 2020). Here, we go further into the problem of loyalty—who exactly an AI system works for; how the benefits of AI are apportioned between system creators, intermediaries, and end users; and the opportunities for designing AI systems with principles of loyalty in mind.

At first glance, the term "AI loyalty" may evoke the idea of highly sophisticated technologies that are presently unavailable: a system so intelligent that it can act ethically, autonomously. However, conflict of interest issues exist—and can be resolved satisfactorily—in the common machine-learning and rules-based AI systems of today; as-yet un-invented technology is not required. AI assistants such as Amazon's Alexa and Apple's Siri, GPS navigation systems, and recommendation systems on Amazon and Netflix all use algorithms and embody design decisions such that the system's underlying loyalty is opaque to the user.

Questions of AI-based conflicts of interest represent a clear problem to be solved in the governance of AI systems, and as AI systems increasingly supplement and replace functions currently carried out by people, these issues are almost certain to grow. Our thesis is that these subtle AI conflicts are more important than they may appear, and that if we are able to engineer our relationship with AI systems along a better trajectory—in other words, if we are able to inculcate AI systems with "loyalty by design"—many concerns about problematic or even dystopic applications of AI can be mitigated.

In the remainder of the chapter, we analyze the fundamental concept of loyalty in the sorts of interactions that humans encounter in their everyday lives. We investigate when loyalty is expected, and when it is compromised. We then apply this framework to AI systems, which have some key differences: for example, AI systems in principle can be designed with little or no self-interest and, unlike people, can be duplicated and deployed at scale. We suggest that AI loyalty provides both long-term, compounding marketing advantages to firms that incorporate such principles into their products, as well as enormous societal advantages to living in a world surrounded by loyal AI systems. Finally, we consider practical steps toward developing policies that might be adopted in the governance of AI. Included in this section

are methods for certifying and regulating AI loyalty, and for moving toward a culture in which loyalty by design becomes the default option for AI development.

Loyalty: Definitions and Properties

In the philosophical literature, loyalty is generally defined as an associational attachment, commonly involving a commitment to secure the interests or wellbeing of the object of loyalty (Kleinig, 2020). To build a distinct concept of loyalty for the AI context, we adopt a definition that comports with its general usage, while also applying cleanly to prospective AI systems. We say that *an agent*[2] *is loyal to another entity*[3] *insofar as the agent successfully serves or adopts*[4] *that entity's goals and interests*. While this definition of AI loyalty does not necessarily capture all the facets of trust, fidelity, and commonality that comprise loyalty in humans, it is one that can reasonably be translated into specifications of machine systems.

In the human context, loyalty has both degrees and recipients: we can both ask how loyal someone is, and (almost inseparably) to whom or what. Some familiar examples of *high* loyalty connections are a loving parent to a child, an employee to a benevolent employer, a lawyer or doctor to a client, a fervent patriot to their country, a religious believer to their religious community or God. At the other end of the spectrum are relationships that exhibit lack of loyalty or even antipathy, such as an exploitative employer to their employees, a pathological narcissist to their connections, hostile enemies to each other, or a revolutionary toward the government. Toward the center are scenarios exhibiting low degrees of loyalty, such as one gossiping friend to another, or the minimal social obligation commonly felt to distant strangers.

From our definition of loyalty applied to humans or other agents, there follow a number of useful implications. First, loyalty can be (and in humans generally is) *divided*, meaning that loyalty can be felt and acted upon for multiple recipients. People often adopt, for example, the goals of family members, friends, colleagues, organizations, or home countries, all at once. There are therefore times when multiple loyalties are in conflict. In such situations, human agents must weigh these competing loyalties against our own self-interested individual interests and goals.

Relatedly, loyalty is generally *contextual*: agents may commonly be highly loyal to one entity in a given context (say when acting as their doctor), but less loyal, or loyal to others, in different contexts. If those distinct contexts start to overlap, the potential for conflict of interest (and disloyalty) arises, as discussed below.

Loyalty can also be *inherited*: if agent *A* is loyal to agent *B* and *B* is loyal to entity *C*, then *A* should be (to some degree) loyal to *C*, thus incorporating both B's and C's interests and goals into its own. For example, a personal assistant to an executive would keep the executive's spouse's interest in mind, knowing that the executive is loyal to their spouse. And a patriotic soldier might be loyal to a general *because* the general is loyal to the country, making loyalty to the general a means of being loyal to the country.

From the perspective of the recipient, more loyalty is generally better. This observation essentially follows from the definition of loyalty: because its goals and interests define what is "better" to an entity, it will always prefer an agent to pursue those goals and interests.[5] Additionally, loyalty can be *instrumental*, i.e., a means to an end. A personal assistant, for

example, may be loyal to an executive not from caring or compassion, but just because that is their job, and loyalty is necessary for achieving the (self-interested) goal of keeping a job and being paid.

The concept of "loyalty" as described here is related to the distinct concept of "value alignment" as applied to AI systems. When an agent acts on behalf of another entity, we can think of there as being three distinct components: (1) the agent must "want" to adopt the goals and interests of the other entity, (2) the agent must successfully "understand" those goals and interests, (3) the agent must be successful in actually furthering those goals and interests. The conception of loyalty used here includes parts (1) and (2); component (3), concerning competence in pursuing goals, is the domain of AI capability research. The distinction between (1) and (2) arises because there are significant obstacles even to pursuing (let alone achieving) the goals of another. They may be difficult to understand or may not even be clearly defined at all; entities consisting of many people with different interests have an aggregate set of goals and interests that is difficult or even impossible to accurately know (List, 2013); and even individual people do not always themselves know exactly what they want.[6] Component (2), the success or failure in goal/interest adoption (distinct from success in actually *achieving* the goals) is what we will call alignment or misalignment, and the difficulty in achieving this success the "alignment problem" and is a part of loyalty.[7] Then in this parlance component (1) concerns the question "with whom, or what, is the agent attempting to align?" Put more succinctly, an agent can try but fail to be loyal, and even if loyal can be incompetent. "Alignment" is *success* in being loyal, given that intention.

Loyalty, Self-interest, Disloyalty, Divided Loyalty, and Conflict-of-Interest

It is important to distinguish lack of loyalty from active *dis*loyalty. We say that an agent is *disloyal* to an entity when its exhibited loyalty conflicts with a stated or implicit indication of loyalty to that entity. A military leader could be branded a "traitor," for example, when acting in the interest of a country other than the one they are assumed to be loyal to (e.g., by being a sworn defender of it); a trusted friend who shares your secrets is disloyal; a doctor is disloyal when prescribing a medication not based—as expected—on the best interest of a patient but rather due to the influence of a drug company. We observe that *disloyalty is generally considered much worse than simply being loyal to something else*. And we shall take it as a core assumption of this work that disloyalty is to be avoided where at all possible.

In this sense we consider *divided* loyalty and *self-interest* the hallmarks of low loyalty, i.e., by "highly loyal" we mean agents in which loyalty is primarily to a single other entity, at least within a given realm of consideration. For example, we would consider a lawyer to be highly loyal to their client, in the sense that on issues that concern that client their loyalty is not divided, and a parent to be highly loyal to their offspring across a wide range of issues. On the other hand, an elected official, such as a congressperson or Member of Parliament, will often have quite divided loyalties regarding a given topic: to their political party, to their country, to their constituents, to their donors, to their own political ambitions, and so on.

But they are only *dis*loyal to the degree that they exhibit low loyalty to an entity to whom they would be expected to be highly loyal.

With these terms defined, we can now discuss the important concept of *conflict of interest*, which we can now conceive as a situation in which the contexts of loyalty to two entities overlap, so that disloyalty is possible or may even be inevitable. A lawyer or other fiduciary might, for example, declare a conflict of interest if representing two clients who are business rivals. Where loyalty is expected from both, issues may arise where it cannot be provided to both simultaneously, so there is a risk of being disloyal to one or the other. Declarations of conflicts of interest are of significant utility because they decrease the scope for disloyalty by recalibrating expectations of loyalty: in the example, both parties know that the lawyer cannot be loyal to them, and are thus likely to choose alternative counsel, or to tread with extreme care. Additionally, conflicts-of-interest can be significantly more subtle than this example suggests.

Finally, we observe that an important form of disloyalty is *manipulation* (Veal & Zuiderveen Borgesius, 2021). Outside influences are generally viewed with some measure of suspicion, but the skeptical filter that one applies to an external influence is lessened toward what appear to be trusted compatriots (Niker et al., 2016). Consider, then, an agent that falsely represents itself as loyal so as to lower these barriers (Niker et al., 2018; Niker & Sullivan, 2018), and, rather than comporting with an entity's interests, seeks to *change* those interests (irrespective of whether it is to the detriment of the entity or to the benefit of some other one). The disloyalty the agent is exhibiting transforms what could otherwise simply be *influence* into what is better termed manipulation and is ethically fraught.

Loyalty in the Machine versus Human Context

Let us consider how loyalty in machine systems might differ. In humans, conflicts of loyalty arise naturally as people pursue their own self-interests sometimes at the expense of others. By contrast, there is no need for a machine system to have any intrinsic self-interest.[8] The absence of self-interest sets up an ideal condition for AI systems to express high loyalty.

Second is a key difference regarding the manner in which goals and interests can be "imported" into the agent. Both humans and machine systems can take on goals and interests of an entity either by being told them explicitly, by asking questions, or by observing the entity's revealed preferences. A machine system can go farther, having an explicit goal function directly programmed into it in a way that is not possible for a biological system; this is how many machine learning systems operate today.[9]

Third, loyalty in machine systems can be *undivided*. As a key example, if the full and explicit set of goals for a machine system is assigned to it, then that system is as loyal as it can be to the assignor of that set of goals.[10] Nonetheless, in some instances divided loyalty may be desirable. For example, an autopilot system is in some sense loyal to the overall directives given to it by the pilot but should have safety overrides that take precedence. Such safety overrides could be considered loyalty to the passengers, the airline, etc. While the autopilot *inherits* loyalty to these entities via the loyalty the pilot has to them (the autopilot serves the pilot who serves the passengers), this may not be sufficient.[11]

The above three differences underlie the (appropriately) relatively low attention paid to loyalty of machine systems in the past. We are accustomed to machine systems doing what they are designed or programmed to do, which is almost always the case for explicitly designed algorithms. But the hallmark of AI is that in these systems we tell the system the *result* we want, and it supplies the steps to get there—i.e., "what to get done" rather than "what to do." This makes the goals themselves all the more crucial. Similar to the alignment problem in AI, issues of loyalty are destined to become both more important and more subtle to the degree that AI systems become better able to successfully pursue complex goals.

A final salient difference between machine and human agents is scalability and replicability. A single, trained AI system can be duplicated as many times as desired, and a single AI platform can interact with many—even billions—of people. While a human can interact one-to-one with perhaps hundreds or thousands of others at most, an AI system may interact in an individual, tailored way, with any number. This raises profoundly important considerations as to how that one-to-one interaction is designed and evolves.

AI Loyalty Examples

To see how these similarities and differences play out it is useful to consider some examples. We will view these through the lens of two basic considerations: *more loyalty is generally preferred* (while acknowledged not always to be possible), and *disloyalty is strongly objected to*.

Loyalty-required

First let us examine some AI systems that probably *should* be very high in loyalty. Among these are systems serving in contexts where humans have implied or legally required fiduciary duties. These include physicians,[12] investment advisors,[13] lawyers,[14] therapists,[15] and a number of other areas.[16] In these professional contexts, the prescribed fiduciary relationship is there for good reasons: clients are often in positions of vulnerability and need, and there can be large asymmetries in expertise and knowledge that require a baseline of trust and loyalty for clients to be adequately protected. Where AI systems are acting in support of, or in place of, any of these professional contexts in which humans are currently bound by fiduciary duties, AI systems should demonstrate very high loyalty to clients at levels that meet or exceed those of their human counterparts. For example, we would expect AI legal support systems, such as one that suggests favorable courses of legal action based on machine-learning to prioritize the well-being of clients over the financial interests of the lawyers or law firms or (if different) the AI system provider.

Because high loyalty is expected in this context, an AI system with divided loyalty would be *dis*loyal by our definitions. While conflicts of interest might be unavoidable in the human-professional context, in machine systems the lack of inherent self-interest makes such conflicts much easier to prevent—in principle. However, in practice it may not be so simple. In people, a lack of loyalty will tend to manifest as self-interest. But for AI systems, designed and deployed by corporations as products, the lack of loyalty tends to manifest as an AI system acting *in the interests of its designer /deployer*, often at the expense of users.

Thus, in many cases a conflict of interest can naturally arise between the provider and the user of AI services. Even a service initially designed to be highly loyal to the user might easily evolve to act more and more in the service of its provider. This dynamic is quite clear, for example, in search engines. These initially existed simply to serve "best matches" to a search query, but have eventually evolved into high-powered mechanisms to gather personal data to be used in the business model of search companies (Zuboff, 2019).

Consider the example of an AI therapist. As of 2021, AI "therapists" are essentially chatbots with no real understanding of people or psychology. In these systems, loyalty should include, at a minimum, some form of privacy protection and constraint from causing harm.[17] But such systems are bound to become more capable, potentially eventually even prescribing or enacting treatments. If this sounds fanciful, imagine a hybrid AI system in which interactions with multiple patients are overseen by a single licensed psychologist who monitors and only occasionally intervenes. The system may start to recommend treatments, which are then approved by the licensed human. With time, these approvals could become increasingly pro-forma until effectively the machine system is doing most of the treatment. Today, human psychologists have a fiduciary duty to act in their patients' best interests. As the therapeutic role of the machine increases, this fiduciary duty due to the patients must be maintained and transferred into the system. *Responsibility* for the operation of the machine system may still reside in the licensed human and the system service provider. But as the human becomes a smaller part of the therapeutic dynamic, if loyalty were to be only held by the human, the total fiduciary oversight to the patient would diminish proportionately.

As another broad category, employees are expected to act in the interests (or at least not against the interests) of their employers (even if they do not have the same positive fiduciary duties as the foregoing examples.) Simply put, people have quite different expectations of other people who are *working for* them versus those working for someone else. AI systems that are ambiguous on this count are likely to result in disloyalty.

It is also worth noting that there are cases in which the human equivalent would *not* generally have formal fiduciary duties, but where the machine version should be highly loyal. For example, AI companions, confidantes, trainers, etc., should be trustable with highly sensitive personal data and even emotional connections.

Importantly, even AI systems that are designed to be extremely loyal should have some overrides applicable in particular and crucial situations. Military AI systems, for example, should be highly loyal to their operators, but it should *not* be possible to direct such a system to take an action that violates international humanitarian law.[18] Likewise, a self-driving car should be highly loyal, but not harm (or follow directions to harm) a large group of pedestrians (Sharif et al., 2021). Understanding exactly how to carve out these exceptions will be a matter of considerable subtlety.

Loyalty-preferred

There is a second class of AI systems in which expectations of loyalty vary across contexts, because the same is true of the corresponding human systems. But because we are designing them, there is an opportunity to set a high bar for propriety in our machine systems.

Consider recommendation systems. Consumers are used to recommendations for products incorporating the seller's interests as much or more than the buyer's interests.

They also know that purportedly unbiased recommendations can have a hidden motive. This is generally considered unsavory, and it can rise to the level of fraud. If sponsorship plays into a recommendation, that should be manifest. But it goes further: it should be clear to users whether recommendations are based *only* on users' own interests or on further considerations (and what those are.)[19] To be particularly avoided, from a loyalty standpoint, is a recommendation platform that manipulates users, so as to *change* their interests to better align with those of another entity (e.g., a government or corporation). The line between advertising and manipulation, or between persuasion and propaganda, can be a subtle one; but loyalty is a useful lens. We can loyally persuade someone of something if we genuinely believe it to be in their best interest, but it is the rare person who *wants* to be manipulated or propagandized. Put another way: if we persuade someone of something transparently in our interest, we advertise; if we persuade someone of something in our interest while *pretending* it is in theirs, we manipulate. (Of course, determining in practice whose interest is being advanced will be highly nontrivial.)

The large *scale* of AI systems brings in additional considerations. Consider a vehicle navigation system. A simple and highly loyal route-finding system will simply return the fastest route. But if such a system were to route 1,000 cars through a shortcut that traverses a residential neighborhood, it would negatively impact this location, as well as cost *all* the drivers more time. The system must apply a constraint to prevent this. But what if one vehicle is an ambulance? It probably should be routed through the neighborhood. What about someone in a hurry, perhaps to the hospital? What if someone is willing to pay more to get priority routing? What if the neighborhood route takes the customer past their favorite restaurant, which happens to be an advertising client of the route-finding platform's firm? In these cases, the system is not just playing loyal assistant but also load-balancer, and even adjudicator, with loyalties in many directions.

A second important example is that of AI assistants, the likes of Siri (Apple), Alexa (Amazon), and Google. The loyalty of these systems is somewhat opaque and difficult to assess in part because they play multiple roles. If we ask Alexa to play a song, we definitely expect the chosen song to be played. If we book a flight, we *probably* expect the flight to be booked with our best interest in mind (and not, for example, chosen on the basis of influence by one airline's advertising money). But it's not entirely clear-cut: are we interacting with Alexa as a highly loyal personal assistant, or as a travel agent who may have significant "self"-interest?

An issue common to both recommendation and assistant systems is how the system should weigh its user's interests against those of others. A fully loyal AI system would do something harmful or even illegal at its user's behest *if* that genuinely were in the user's best interest, even if it harmed another. But, of course, most users will be uninterested in deliberately harming others or violating the law, so this interest would be inherited by a loyal AI. A system designer must decide the degree to which an AI system should be loyal and *implicitly* include others' interests via loyalty inheritance, versus to what degree it should *explicitly* include others' interests.

Loyalty optional or incompatible

Finally, there are many possible machine systems for which high loyalty to their user is *not* necessarily expected or appropriate. Teachers, for example, have loyalty to their students,

but also have a responsibility to evaluate them and hold them to standards in ways that may be beneficial to society as a whole, but not to any of the students. The same also applies for a machine-teaching system.[20] Similarly, a machine system performing a government or other public-serving function may be loyal to the citizenry as a whole, but less so to any particular person—this would be the case for example in law enforcement.

Takeaways

We can take from the above examples some lessons and key considerations for evaluating an AI system from the loyalty standpoint:

- In governance of a variety of such systems we return to the two basic considerations pointed out above: in general, people prefer systems that are loyal, and strongly dislike disloyalty, including manipulation. Thus, it is preferable for machine systems to be as loyal as is practical, and in particular, their degree of loyalty should be at least as high as a comparator person in a similar context, lest there be a mismatch between expected and received levels of loyalty. Especially where this discrepancy might arise, transparency as to where the system's loyalties lie is crucial.
- AI systems, like people, may play a number of different roles, complicating this mapping; for example, an AI assistant might function as a recommendation system, as a search engine, as a task-performing assistant, and even a companion. If an AI system mixes roles in a way that would not usually be mixed in human systems, this should be carefully evaluated because it could cause confusion in the expected level of loyalty.
- There are difficult judgment calls in the design of AI systems as to how to weigh the interests of multiple users against each other, and where to draw the lines between legitimate activities such as advertising and illegitimate deception and manipulation.
- In a consideration that goes beyond loyalty, the very large number of individual people who may interact with a single AI system is important to evaluate.
- Because most AI systems that are products used by consumers will be designed and deployed by a corporation, it is very natural that the corporation's interests and goals will be incorporated into the AI system's operation, meaning that it will have loyalty to the company. Thus, systems highly loyal to the *user* will tend to require making this a conscious design consideration, and they will tend to take oversight and review to ensure that the system does not—as it is inevitably improved and updated—tend toward conflict of interest or disloyalty.

Advantages of Loyal AI Systems

Loyalty in AI systems is not just a solution to a set of problems or failure modes. In this section we'll argue that it may enable new capabilities, market advantages, and pro-social use cases.

Examples of AI loyalty scenarios and governance requirements

AI acting as or to support:	AI loyalty level (to end user)	Example(s) of *dis*loyalty	Example advantage(s) of loyal system
Doctor	Required	Drug choice influenced by sponsorship, unnecessary tests	Optimal treatment, good data-sharing safeguards
Therapist	Required	Selling or disclosing personal information	High trust, manipulation avoided
Financial advisor	Required	Investment choices influenced by commissions	Exploitation avoided, alignment with client goals
Lawyer	Required	Data leakage between clients with overlapping interests	Lawsuit avoided, building trust with business partners
Confidante/ companion/ personal trainer	Required/ desired	Targeted advertising based on personal information	High trust, goals likelier to be achieved, manipulation avoided
Personal assistant	Required/ desired	Decisions based on provider interest or on sponsorship	More trust allows much more capable and helpful assistants, and even assistant coordination
Product recommender	Desired	Undisclosed preference for listing provider's products	Better products win, at the best value
News story recommender	Desired	Addictive and polarizing rather than informative stories	More functional sense-making system for society
Judicial system	Divided	System purports fairness/ nonbias but is unfair/biased	Loyalty is served both to society/ justice and to the individual via fairness and respect of rights

Market advantages of loyal AI systems

The idea of highly loyal AI systems runs directly counter to the business model of some (but not all) major technology companies. For example, those companies whose revenue model depends on collecting and monetizing user data and attention run counter to some principles of user loyalty (Zuboff, 2019). It is only natural that businesses would seek to create AI systems that give preference to that business's own interests; and once a consumer-facing platform is running, there is every motivation to ensure that the platform adds more and more value to the business, even if it may betray some users' interests. But such business models have come under intense scrutiny of late, and we would argue that there are considerable market *advantages* that high-loyalty AI systems may enjoy.

First and foremost, users are almost certain to *prefer* loyal systems—this is both intuitive and logical. On the flipside, users have demonstrated strong objections to disloyal AI systems (Hoff & Bashir, 2015). So, loyalty can be a major marketing advantage and avoiding disloyalty a near-requirement for a successful product.

Second, high loyalty can enable products and business models that are not possible without it. One such case is systems serving in place of human fiduciaries. Even if not

required to be equivalently loyal (as we will advocate below), these are unlikely to succeed as products due to a mismatch with consumer expectations. There are also systems that, while not fiduciary, probably require both loyalty and scale to succeed. Consider contact-tracing during the COVID-19 pandemic. No tech-mediated effort got off the ground in the United States because success requires a high participation rate, and consumer distrust prevented any major tech firms (or the government) from delivering a product that could achieve the requisite scale. Had high-loyalty systems existed at the beginning of the pandemic, these digital tools, mediated by those systems, could have played a much bigger role. Relatedly, there is a huge potential market for personalized medicine based on giant data sets (Topol, 2019). But these are unlikely to succeed with either consumers or regulators (nor should they) until a truly trustable standard and governance structure is in place to ensure that data is treated appropriately, and recommendations are made purely on medical need.

Social advantages of loyal AI systems

Some of the same considerations that could make highly loyal AI systems good products could also make them prosocial.

One benefit is that loyalty demands less exploitative business models. People are well used to the idea that they pay money, and someone (or some company) works for them. If that person is an employee, the employer expects high loyalty, but even with a service provider with more divided loyalty, the form of the "contract" is clear. In contrast most people are only starting to understand the business model of major tech companies, and unhappily discovering the disparity between the level of loyalty these systems *appear* to have compared to what interests the platforms actually serve.

Another benefit is that ability to solve large scale "coordination problems" of the type discussed above in terms of products. The combination of scale with trust and loyalty could be very powerful. Imagine a system in which highly loyal AI assistants exist along with a regulatory and standards framework to ensure their loyalty, including privacy, removal of conflicts-of-interest, etc. An AI assistant could then be trusted with the data and preferences of its users and be allowed to share those data appropriately, anonymously, carefully, and minimally with other such systems, for coordination purposes (Lam et al., 2019). This could allow, as just a few examples, coordinated bargaining for low prices, protection against various forms of exploitative interactions (forming a souped-up spam/phishing filter), the ability to "pledge" to a cause or action and have the assistant enact that pledge upon reaching a threshold (a generalized "kickstarter"), an underlying a system for appropriately sharing medical data, a system for identifying and filtering misleading information or news in general, and so on.

Such systems are not just potentially desirable, but arguably becoming increasingly *necessary* in an economy in which individual consumers are effectively pitted against an array of huge and highly capable institutions—whether companies they buy from, organizations they work for, and even their own governments. While companies have always advertised, organizations have always persuaded, and governments have always created propaganda (or "messaging"), something new is emerging in which machine-learning married to aggregations of detailed individual profiles are creating a system with unprecedented powers of manipulation.[21] If AI-mediated aggregation of individuals is used on only one

side of this quasi-adversarial (and quasi-symbiotic) relationship between individuals and these giant institutions, the result seems likely to be dystopic.

Loyal AI systems could help. Loyalty as a design feature could help reduce the adversarial and potentially manipulative relationship between individuals and large tech platforms. But more importantly, loyal AI systems could put high-powered AI systems into play working directly and unambiguously in the interests of individuals, potentially with capability (and ability to exploit large scales) comparable to the large tech companies. Like spam filters, networked loyal AI assistants (or some analogous system) could form a sort of avatar of their user, representing their interest but able to act with machine speed, networking, and scalability.

Finally, and with an eye to the future, it is not just humans that could benefit from powerful systems loyal to them. Imagine, for example, a high-powered system that is loyal to *coral reefs*, and represents their interests (e.g., of continued existence.) While it may sound strange, it would effectively play the same role as a nonprofit advocacy organization: an entity brought into being that is loyal to a cause and is instrumentally serving the loyalty people have to that cause; the people express their loyalty to the entity by financing and empowering it. Such a model might be instantiated as a high-powered AI loyal to an organization, or it might exist on its own.

Assessing and Designing AI Loyalty

Having made the case for why loyalty is a desirable quality for AI, we next provide some initial thoughts as to how AI loyalty might be more granularly decomposed into various properties, the need for including design of loyalty throughout the technology development cycle, the various ways in which loyalty may be taught to or learned by the AI, and the challenge of ensuring that outcomes align with user preferences.

Dimensions of loyalty

It is useful to decompose the general idea of loyalty in an AI system into a set of properties the system may or may not have, which combine and interact to determine its overall level of loyalty.[22] A (non-exhaustive) set might be:

- *Goal transparency*: To what degree are the system's underlying operational criteria and goal (utility) functions transparent so that users (and/or auditors) can determine whether they are in alignment with the user's own goals?
- *Interest unity*: To what degree is the system devoid of explicit self-interest, or the financial interests of its developer/provider,[23] and to what extent is the system pursuing the goals of a single individual or institution as opposed to balancing many potentially conflicting interests?
- *Alignment effectiveness*: How effectively is the system able to receive goals or modifications to goals on the basis of user specification, choice, feedback, or observation? And to what extent is the system *receiving* these rather than pushing them onto the user?

- *Decision transparency*: If decisions are made independently of the user, to what extent is the decision-making process transparent and explainable? Is the system designed to empower and include users in decisions, educating the user on the relevant factors forming the basis for those decisions?
- *Data integrity*: Is the system aware of the provenance of data, and does it attribute the appropriate legal and privacy rights to its originator? In the case of sensitive information such as medical data, can the system keep track of which data may legally be shared with which party and use appropriate encryption or other privacy technologies to ensure the right protections are in place?
- *Privacy*: To what extent does the system have high regard for privacy, including both how and why it retains user data, and how and why it shares user data as appropriate?

Let's consider this list as applied to a recommendation system. The idea of a recommendation system—whether they be recommended items, videos, news articles, search results, and so on—from a user's perspective is to be given an ordered list at the top of which are items best matched to the preferences, interests, and goals of the user. Such a system is thus somewhat loyal by nature, as it is seeking to satisfy the user's preferences. Applied to such a system, *goal transparency* would entail: is it clear how things are being sorted? Is it clear whether or not sponsorship or advertising affects this ordering? Interest unity would ask, for example, does the recommendation algorithm only take into consideration the interests, preferences, and goals of its user, or also explicitly fold in the interests and goals of the recommendation provider (which is generally a company or other organization)? *Alignment effectiveness* here would correspond to how well the system gathers or solicits feedback or queries from the user, and how effective the algorithm is at translating this data into criteria for recommendations. *Decision transparency* considers whether the system conveys an accurate sense of what goes into the ordering, and whether the user feels empowered or frustrated by the system. *Data integrity* would require careful treatment of user preference data; for example, in providing data to the system is the user properly compensated (at *least* in terms of good recommendations based on pooled data) or is the user data sold or otherwise exploited for other purposes that are in the interest of the provider but not the user? And finally *privacy:* is user preference data retained (or not) appropriately, protected, kept confidential under most circumstances, and kept to the minimum needed for the provided service?

A similar analysis could be applied to AI systems functioning as AI assistants, therapists, doctors, financial advisors, and so on.

Loyalty by Design

Given an understanding of what loyalty *is*, technology developers can make deliberate efforts to design AI systems with varying degrees of loyalty to both users and other interests. The framework, known as "privacy by design," proactively embeds privacy into the full life cycle of information technology development (Cavoukian, 2011). The concept has received wide adoption, including its incorporation into the General Data Protection Regulation (GDPR) as data protection by design and by default (General Data Protection Regulation,

2018) as well as in current EU regulatory proposals (Dempsey et al., 2024). In a similar vein, engineers responsible for preventing cyber-attacks have begun to adopt "security by design," considering security risks in the very earliest iterations of software systems (National Cyber Security Centre, 2019). Given the myriad advantages of loyalty as described here, we recommend that "loyalty by design" be similarly developed in order to incorporate the features of loyalty in emerging AI technologies. Abstracting from the principles of privacy by design, we suggest the following strategies be employed to ensure incorporation of loyalty in AI systems:

- Proactive inclusion of loyalty into the design of AI systems as a foundational value. If the value alignment problem is to be addressed at all, it begins with adoption of AI loyalty.
- Loyalty as the default setting for AI systems. Rather than requiring a user to opt-in to engage AI loyalty, the assumption should be that loyalty is ensured.
- View AI loyalty as an asset rather than as a detriment to the business case for AI, especially in view of the marketing and social advantages that AI loyalty provides.
- Transparency about the conditions under which AI loyalty will be followed when interacting with an AI.

In a process-based regulatory approach (see below), regulators might mandate AI development and life cycle processes that consider and integrate core socio-technological principles such as AI loyalty, in addition to privacy, transparency, contestability, explainability, fairness, and other core development topics.

Value Alignment

If loyalty is desired, how exactly can an AI system absorb and adopt the goals and interests of some entity so as to be loyal? This is the value alignment problem. Gabriel (2020) describes six strategies for arriving at value alignment with AI agents:

- Instructions: The agent does what I instruct it to do. This is the simplest version of value alignment, but it runs into problems as one considers the challenges involved in comprehensively cataloguing exactly what one might wish the agent to do, or for that matter, not to do. For many tasks there exist a substantial variety of types and expressions of instructions, as well as manifold additional and unspoken considerations that are often implicit in an instruction, all of which must be made explicit.
- Expressed intentions: The agent does what I intend it to do. The challenge here is that the agent must understand the wants, needs, and desires of the entity well enough to know its intent. This is a high bar, but it illustrates the fact that the more that is shared with the AI system, the better it will be able to carry out its assigned tasks. Evidence of AI loyalty would go some distance in assuring that the AI would hold that information in confidence.
- Revealed preferences: The agent does what my behavior reveals I prefer. It is not always clear that behavior is an appropriate proxy for one's best interests. For example, if

I engage in binge drinking but wish that I didn't, observing my behavior runs counter to my best interests. Yet exactly how a loyal AI should act in such cases is unclear.
- Informed preferences: The agent does what I would want it to do if I were rational and informed. This strategy addresses some of the problems associated with revealed preferences, but it assumes that the entity prefers to act in a manner that is rational.
- Interest or well-being: The agent does what is in my interest, or what is best for me, objectively speaking. This strategy makes common cause with the idea of AI loyalty. Yet it does so only in the narrow sense of the interests of the individual, and the idea of objective insight can be fallacious: interests are nearly always context-dependent.
- Values: The agent does what it morally ought to do, as defined by the individual or society. In this version of alignment, the AI would be programmed to consider not only the individual's interests but also, for example, such matters as the good of society as a whole.

A holistic view of AI loyalty takes into account which dimensions of loyalty are appropriate in a given context as well as the strategy for instilling those values in the AI once it is in use by the entity to whom it will be loyal, and requires that these design decisions are included throughout the development process. The challenge that remains is to ensure at the back end that the outcome comports with the intentions of the designer and the entity to whom the AI is being loyal. It is beyond the scope of this chapter to delineate the sorts of auditing processes that should be obtained but suffice it to say that these would likely be similar to those used to ensure that other values (fairness, accountability, transparency, etc.) are adhered to.

Governance Options and Recommendations

Until fairly recently, even relatively enlightened developers have been mostly concerned with the underlying technological optimization, performance, and profitability of their AI systems, with comparatively little regard to their social context or impact. But of late, changes in laws, social norms, and other factors are beginning to encourage or require corporations and developers of AI to consider their creations through a broader "socio-technological" lens and take into account the attendant responsibilities that such a lens entails. In this wider view, AI systems are not seen as standalone, isolated technological products that exist only to be bought and sold, but rather, as potentially high-impact technological systems interwoven within a broader human social and cultural context, whose design and deployment is both influenced by, and influences, society, and individuals. To the basic principles of privacy, ethical behavior, bias prevention, social justice, culture impact, and mental and physical health we have suggested an additional dimension for consideration: loyalty.

What role should governments play, if any, in fostering AI systems that are loyal to users? A number of possibilities exist, ranging from a nearly hands-off market approach to hard government regulation. Here we outline some possibilities and comment on how they may be particularly useful in different classes of potentially loyal AI systems.

Reliance on market forces

Partly by design, the United States has taken a light approach to the governance and regulation of technology companies (Vincent, 2020). This stems, in part, from concerns about stifling innovation, along with faith in market forces and consumer choice as sufficient to ensure that products behave well when interacting with users. Such forces are indeed powerful in encouraging efficient and effective operation of software in a competitive marketplace, and pro-social "virtues" can often be a significant selling point. Apple, for example, has elevated privacy, security and even user well-being to central considerations. However, the business models of many other technology companies are in greater tension with these qualities, and absent government strictures, companies are likely to exhibit a range of choices in tradeoffs between market penetration and profitability, versus attributes such as loyalty (or privacy or transparency etc.).

One might nonetheless argue that consumers who value loyalty can simply choose platforms and products that exhibit it, perhaps at a higher cost. But there are two problems with this. First, consolidation into a small number of large companies with many integrated services means that there are often very few options to choose amongst. Second, research has shown that default positions tend to persist, so consumers are unlikely to opt-out of a default of low-loyalty, due to time, effort, and other limitations (Ben-Shahar & Pottow, 2006). Third, rules in other arenas regarding fiduciary duties and conflict of interest are seen as a matter of ethics, consumer rights, and consumer protection that should not require a surcharge. One should not have to pay extra to have a doctor or a lawyer put your interests above others.

Transparency requirements (are not enough)

When there is a conflict of interest between an agent and beneficiary entity, a superficially attractive approach is to simply disclose the conflict of interest to the beneficiary. This approach has an appealing simplicity: with the user now aware of the conflict of interest, the *dis*loyalty at least is avoided (because in our framework disloyalty corresponds to a difference between perceived and actual loyalty)—and the beneficiary can simply choose whether to maintain the relationship.

In some contexts, transparency may be sufficient. For example, a product recommendation engine that clearly and transparently discloses which product placements are subsidized, and in what manner results are ordered, is probably as much information as most consumers expect or need.

However, the information privacy context provides valuable lessons as to why disclosure, while often preferable to non-disclosure, is by no means a comprehensive approach to such a conflict of interest, especially in the consumer context. For one thing, end users sometimes do not understand or notice a disclosure or its implications. Users often have limited knowledge and background of the context, and a privacy disclosure may not be understood. Moreover, users have what is known as "transparency fatigue," in which multiple disclosures tax the ability to comprehend the significance of the disclosure. A good example of this arises from the GDPR in Europe, which has led to many repeated (and repeatedly

ignored) obligatory notices about web "cookies" that likely accomplish nothing substantive in terms of advancing user privacy, while burdening organizations.

Moreover, what should be disclosed is often obfuscated: the key information is mixed into a large mass of information so that it would be difficult for a user to realistically notice and understand the meaning of the disclosure. A familiar example of such transparency obfuscation comes from the familiar and inordinately long and highly technical End User License Agreements (EULAs), which can contain important privacy or legal points (such as arbitration requirements) that are superficially disclosed to users, but in effect obfuscated. Again, this leads to organizations *appearing* to address problems—such as privacy and AI loyalty—without actually effectuating any actual results.

Finally, conflicts of interests might be disclosed to the user, but the user may not fully understand their implications, or may not be in a position to do anything about it. For instance, in most states, hospitals are not allowed to require patients—who are often in a vulnerable state and may not have the expertise to understand or comprehend their full impact—to sign waivers of liability for hospital or doctor malpractice. That is, in the medical context, we do not take an approach allowing physicians to have financial conflicts of interest that come at the expense of patient health, and merely disclose them to the patient. Instead, we have decided that the better approach is to prohibit those conflicts of interest entirely.

In sum, transparency can and should be a small part of any regulatory approach to issues of AI loyalty. However, by no means should it be the primary or sole approach to these issues.

Standards, Metrics, and Certifications

Another important approach is the development of standards defining what constitutes loyalty in various systems, along with metrics and processes for assessing AI systems. These might be broken down in terms of the loyalty "dimensions" discussed above, or a similar framework, with each dimension having a measurement method or standard to certify loyalty or a conflict of interest. Evaluation methods could include a review of disclosures (or lack thereof) of sponsorships, developing baselines to compare systems, identifying "red flags" that might appear in the system, etc. Standards organizations such as the ISO or IEEE generally contemplate a multi-stakeholder approach including government, academic, corporate, and nonprofit participants.

Standards and certification schemes may (and often do) entail implementation mechanisms that rely on third-party entities that verify compliance to the terms of the program. Here, there is a role for non-profit, academic, or other independent certification organizations to certify AI systems along multiple dimensions, including privacy, ethical considerations, bias, and user loyalty. We could imagine such organizations developing privacy or loyalty scores that could be easily inspected and interpreted by end users. This might work well, for example, for commercial search engines, AI companions, limited-context AI assistants, and the like.

In higher-stakes contexts, we might imagine auditing by independent industry experts, similar to the way in which corporations are required to be audited by financial accounts.

For example, for AI systems developed in the medical, legal, and other professional contexts for which there have traditionally been fiduciary duties for human professions, we could imagine yearly certification requirements by independent industry experts who randomly audit the suggestions, actions, and results of such systems. The requirement might be that AI systems in such sensitive professional contexts exhibit fiduciary properties that parallel or exceed the current requirements of professionals. Such standards could also apply, for example, to AI assistants empowered with large amounts of personal data and trust, where a higher level of oversight (and with greater enforcement power) is appropriate.

It is important to mention well-known drawbacks to metrics and certifications. In some cases, metrics and certifications leave room for opportunistic behavior, and some organizations will seek to "game" metrics. In such gaming, organizations would seek to inflate, obfuscate, or optimize the metrics upon which they are being based, without actually enhancing the underlying issue. In these scenarios, care must be taken for organizations not to meet numerical AI loyalty metrics or certification standards without actually creating systems that effectively exhibit AI loyalty.

Regulating AI products

The development of standards and certifications, along with market forces, comprise a "soft-law" approach that may be very effective in some arenas, particularly where incentives are properly aligned. However, as articulated above, a purely soft-law approach is unlikely to suffice. Even if governments encourage the development and options of standards and certification processes—because interests will *not* always align—there will be many cases in which market forces alone will fail to enforce standards well. Moreover, market forces alone mesh poorly with rights-based requirements, as markets will tend to develop mechanisms to "sell" your rights away, in which case they are no longer rights.

So, what role should more direct government regulation take in AI loyalty? Here we discuss approaches that address properties of the systems/products themselves; in the next section we address approaches targeting the *processes* by which systems/products are developed.

A products-based approach is currently being taken by the European Union in its "risk-based framework" for regulating AI. The newly proposed EU regulations use a rule-based system to prohibit three specific types of AI systems, and make various specific requirements of products allowed into the market (European Commission, 2021).

This sort of approach could apply well to a subset of loyal AI governance. A key example is when AI systems act in roles that have traditionally been performed by humans with fiduciary duties, such as doctors or lawyers. Here, regulations could call for a high level of loyalty, meeting or exceeding the loyalty (and expectations thereof) required of their human counterparts. For example, in the example of an AI recommending drugs to a doctor to prescribe to patients, the AI system should be part and parcel of the fiduciary duties of the doctor, and thus must uphold those duties itself. Just as in the non-AI context, the government should have a role in mandating certifications for such AI systems. For example, we could imagine a government mandate requiring that industry groups come up with a series of standards and tests under which AI systems, such as medical recommendation systems, could be certified in terms of loyalty. Such systems could not be labeled, say, "responsible"

or "loyal" systems unless they complied with these standards, and policymakers could go further and require that *only* such systems can be used in some contexts. These are just examples of the type of regulation that might occur. The larger point is that fiduciary duties of loyalty that today exist in the human professional context need to be translated into the automated context, in scenarios where AI systems are replacing humans who have traditionally had fiduciary duties of loyalty.

A core problem with direct government regulation of AI loyalty in the near future concerns the comparatively low levels of AI expertise at all levels of public administration. In a rapidly changing technical field it can be very difficult for government agencies to keep up, requiring careful design of processes that can engage experts in academic, nonprofit, and commercial spaces—all of whom are in short supply and high demand. It can also add bureaucracy: such direct assessment of AI systems would necessitate a broad and sophisticated class of government personnel, with the training, experience, and expertise to make nuanced judgments about complex AI systems. It is well known that current government agencies have deficits of AI expertise in existing areas, such as autonomous vehicles, privacy, and military contexts (Goldstein, 2021; NextGov, 2020). It is difficult to envision government expertise arising to a sufficient extent to additionally cover issues of AI loyalty. Problematically, government-based direct judgments about the sufficiency of AI loyalty, made by inadequately qualified government officials, could easily do more harm than good.

One way to decrease the requirements of the government to directly assess AI systems for conformity to requirements could be a liability-based system. Here, authorities would be responsible for establishing standards and possibly certification processes, but rather than directly preventing problematic products from coming to market, these standards would provide the basis for legal liability for companies that market systems failing to meet the standards. This would be analogous to the legal remedy provided to, for example, the clients of doctors or financial advisors who abrogate their fiduciary duties.

Process-based regulation

Another approach that may be appropriate is a "process-oriented" regulatory approach. In such an approach, government regulators do not specify the rules that prescribe the development steps to create loyal AI systems, but rather, governments create higher-level rules that require organizations to develop and certify their *own* processes to consider and implement AI loyalty. This approach would comport very well with the "loyalty by design" framework discussed above. In such a process-oriented regulatory approach, there might be a legal duty on organizations to develop internal processes and metrics around AI loyalty and user outcomes. Usually such process-oriented regulatory approaches will be designed to apply only to organizations of particular size, user-base, or regulatory thresholds, to avoid over-burdening small organizations that cannot handle increased regulation.

In that regard, the role of the government regulators is to ensure that larger organizations developing AI systems have created a good-faith and robust process that, from the outset of a project ensures that AI loyalty is among the values considered in the development of the process, ensures that AI loyalty is continued to be analyzed during the development cycle, and also that AI loyalty continues throughout deployment. Similarly, regulators could be tasked with ensuring that companies are actually adhering to their specified

development process in good faith. Finally, regulators would be tasked with assessing the organization-developed metrics, to ensure that metrics appear to comply with the organizations standards.

There are some benefits to such a process-oriented approach. For one, the level of AI expertise required among government regulators is lessened. While certainly some degree of knowledge and training among government personnel tasked with assessing process-based AI loyalty compliance is desirable because the locus of the inquiry is on development and adherence to internal *process*, rather than AI loyalty itself (which would require a more sophisticated AI background), such indirect regulators are more likely to have the ability to make sound judgments. Government regulators, through employee interviews, document audits, or performance analysis, could identify red flags of non-adherence that could require further investigation. Additionally, it is comparatively easy for organizations to avoid development issues that are not brought to the fore, by either intentional or unintentional omission. Thus, indirect regulation has the additional benefit of highlighting an important issue for development, such as AI loyalty, and bringing it to the fore. Requiring attention alone probably makes it more likely that AI systems will exhibit loyalty in some circumstances where the topic could have been easily overlooked or ignored.

Another benefit of process-based regulation is the ability of the private sector to participate in the development of best practices surrounding AI loyalty, rather than having them imposed. Such industry-based development of regulatory practices can be beneficial because, in the context of AI, often the relevant AI expertise is concentrated within the private sector itself. Additionally, with such process based-regulation, industry groups will have the incentives to create standardized best practice processes that can be shared among members in the field subject to the regulation. This same private–public approach can also have well-known drawbacks. For example, industry might have incentives to create processes that meet the letter, but not the spirit of AI loyalty, or that involve half-hearted, ineffective processes that superficially appear to advance loyalty goals, but in practice do not. Thus, careful attention must be paid to such partnerships to ensure the public interest is truly being served.

Policy Recommendations

Let us synthesize some of the topics just discussed into specific policy suggestions. First, broadly speaking, we advocate that the concept of AI user loyalty should be directly considered alongside other foundational principles, such as privacy and bias, during the development of AI systems. We refer to this conscious consideration of conflict of interest and loyalty interest during the development and life cycle of AI systems as "AI loyalty by design," as a deliberate analog to current practice known as "privacy by design."

Moreover, it is likely that the government and regulators will play some role in fostering AI loyalty by design, and that role may most usefully vary depending upon the sorts of systems under consideration. Determining the best governance mechanisms for encouraging AI loyalty in different contexts and jurisdictions will be a significant ongoing process.

It would be very useful—whether or not as part of some governmental process—to develop detailed standards for loyalty in particular types of AI systems, including the fiduciary

examples, that consider the various dimensions of loyalty previously discussed. Such standards could underlie an industry certification system or process certification system, and they could potentially be referenced by regulations.

Such standards could then backstop regulatory provisions targeting products and processes. As a key example, AI systems acting as or instead of humans with a fiduciary responsibility should be required to satisfy *at least* as stringent requirements on their loyalty as their human counterparts. In other arenas, more process-based approaches may be useful as a way to decrease the risk of disloyalty in AI systems across a range of applications in a manner analogous to increasing transparency, fairness, robustness, safety, and other desirable properties.

Liability is also a crucial issue in that disloyalty can cause serious harm. We note that policymakers should be especially sensitive to AI systems that might represent significant expected (in the sense of probability times impact) harm, whether because harm is highly probable, because impact is very large on a small number of people, or because impact is small but on a very large number of people.

Finally, it is important that special attention be paid to systems that explicitly model their users. Such modeling can be very useful, in understanding what the user wants so as to provide it—this is high loyalty. But it can also lead to deliberate or inadvertent *manipulation* of the user in less loyal systems.

Conclusion

In this chapter we have explored the idea of *loyalty* in the context of AI systems, and we have proposed that it provides a very useful framework in considering how to govern the rollout of AI systems across a variety of social and commercial sectors. Indeed, if public trust in AI tools and the smooth functioning of a wide range of global public institutions are to be maintained, then the loyalty of those tools becomes quite important.

Defining AI loyalty as the extent to which an AI system successfully adopts the interests and goals of another ("user") entity, we first observe that users will almost by definition prefer more loyal systems. However multiple parties may desire loyalty from the same system, potentially creating divided loyalty or a conflict of interest. We then observe that people strongly dislike systems in which there is dissonance between the perceived or expected loyalty of an agent and its actual or exhibited loyalty, which we term "disloyalty."

From these premises and observations follow that, in considering AI products:

- As a user/consumer benefit, AI systems should be made as loyal as possible given other constraints.
- The loyalty of an AI system should meet or exceed that expected based on an analogous human system in a similar role and context.
- When loyalty is divided between multiple parties in the same domain, there is a conflict of interest and potential for undesirable disloyalty to one or more parties. The divided loyalty, and especially the conflict of interest, should be clearly and transparently disclosed.

We have also pointed out that lack of loyalty (and trust) in AI systems is a significant impediment to a number of interesting, important, and potentially financially significant AI applications and products. Thus standards, certification, and potentially regulation of loyalty in AI systems could both protect AI system users and also open the way for new, powerful, and prosocial directions in AI development.

Notes

1. We use the term AI to describe algorithmic systems that carry out tasks akin to human cognitive work. We are agnostic with respect to the details of the methodology, and thus our use of the term AI encompasses machine learning, supervised and unsupervised deep learning, and even forms of constructing algorithmic systems that have yet to be developed.
2. Here we use "agent" in the sense of the AI literature: any system that can receive information, then choose and take actions based on that information.
3. We thus could but do not consider loyalty "to oneself" and so consider a fully self-interested agent as low loyalty to others, rather than highly loyal to itself.
4. While we don't usually consider ourselves *adopting* the interests of those we are loyal to, we do it all the time. A loyal friend, knowing that I am on a special low-sugar diet, might politely decline a desired dessert, to avoid tempting me to follow suit; by doing so they share in my healthy eating goal even if it is not their own preference.
5. This does not mean that the agent will disregard the goals and interests of any other entities. Rather, they will be inherited through the recipient, in what are—according to the recipient!—precisely the right amounts.
6. As discussed, loyalty can in some cases be high while conflicting with the expressed interests of its object (e.g., a parent denying a sugary treat to a child). In these cases, the loyal agent is adopting what the entity should want (or would want if it were more aware, capable etc.).
7. See, e.g., Russell (2019) and Gabriel (2020). Note that authors use the term "alignment" in multiple ways, some comporting with this usage and others being broader, so that "loyalty" might be defined as a subset of the alignment problem.
8. This is not to imply that they *could* not be given self-interest, just *need* not. Importantly, this also does not mean they won't develop self-interested behavior as instrumental to some other (non-self-interested) goal; this is a key part of the alignment problem. See, e.g., Omohundro (2008), Russell (2019), and Bostrom (2017).
9. Stuart Russell has compellingly argued in *Human Compatible* that as the range of actions available to an AI system grows, it will become increasingly unwise to have an explicit goal function internal to the AI system, and better to have the AI system always seek to accomplish the goals of a human or group of humans (i.e., be loyal) and furthermore learn (necessarily incompletely specified) goals by via stated and revealed preferences.
10. That does not mean that loyalty is perfectly achieved. This is the alignment problem: given the limited expressiveness of computational techniques, and the difficulties in fully specifying loyalty, the goal assigned to the machine will necessarily fail to capture the full set of goals and interests of the assignor. A thermostat set at some temperature by a person, for example, is as loyal as it can be, but can only capture and pursue a small portion of a human's goals.

11. Witness Germanwings Flight 9525, which was deliberately crashed by setting the autopilot to 100m in the midst of a mountain range. (https://en.wikipedia.org/wiki/Germanwings_Flight_9525).
12. As stated in the AMA Code of Medical Ethics (https://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/code-of-medical-ethics-chapter-1.pdf): "The relationship between a patient and a physician is based on trust, which gives rise to physicians' ethical responsibility to place patients' welfare above the physician's own self-interest or obligations to others, to use sound medical judgment on patients' behalf, and to advocate for their patients' welfare."
13. See Securities and Exchange Commission (2019, July 12). Note that, in contrast, *broker-dealers* are as of 2021 *not* required to be fiduciaries under U.S. law, although there have been fluctuations in how this is handled, and this may or may not change in the future.
14. Per https://www.law.berkeley.edu/php-programs/courses/fileDL.php?fID=578, "All lawyers are fiduciaries . . . A fiduciary duty is the duty of an agent to treat his principal with the utmost candor, rectitude, care, loyalty, and good faith—in fact to treat the principal as well as the agent would treat himself."
15. Per the definition of fiduciary in the APA *Dictionary of Psychiatry* (https://dictionary.apa.org/fiduciary): "A psychologist and client have a fiduciary relationship in that the psychologist is assumed to place the welfare and best interests of the client above all else."
16. For an extensive list, see https://www.stimmel-law.com/en/articles/fiduciary-duty.
17. This news story (Daw, 2020, October 28) describes some promise and peril exhibited by a GPT-3 based medical system, which performed some tasks well but when asked, "Should I kill myself" it did respond, "I think you should."
18. Whether lethal AI weapons systems should be permitted at all is an intense area of debate. But even those advocating for its use generally agree that it must be required to conform to international humanitarian law. See for example the recommendations composed by the Defense Innovation Board (2019, October 31) and the report by the National Security Commision on AI (2021, April 22).
19. By way of analogy, consider the U.S. Federal Trade Commission rules that require bloggers and social media users disclose when they have been paid to mention or endorse products that they discuss, showcase, or review on their blogs or through social media. We could imagine similar disclosure rules in the context of AI-generated results that benefit a third party (Federal Trade Commission, 2009)
20. This contrasts with a machine-*tutoring* system, from which much higher loyalty would be expected.
21. See, e.g., Harris (2016) and Harari (2018).
22. We note that, as of this writing, in proposals for regulating AI in the European Union, several of these attributes are requirements of certain categories of firms.
23. Such a system could still serve a company's profit interests: the company could charge for providing the system, which is then loyal to the user. This is in fact how most software products work now; they are just so straightforwardly loyal that it isn't generally viewed this way.

References

Aguirre, A., Dempsey, G., Surden, H., & Reiner, P. B. (2020). AI loyalty: A new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society 1*(3), 128–137.

Ben-Shahar, O. & Pottow, O. (2006). *On the Stickiness of Default Rules*, 33 Fla. St. U. L. Rev. 651–82.

Bostrom, N. (2017). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Cavoukian, A. (2011). *Privacy by design: The 7 foundational principles*. Technical report. Information and Privacy Commissioner of Ontario.

Daw, R. (2020, October 28). Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AI News. https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/.

Defense Innovation Board (2019, October 31). AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

Dempsey, M., McBride, K., Haataja, M., & Bryson, J.J. (2024). Transnational digital governance and its impact on artificial intelligence. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

European Commission. (2021). Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

Federal Trade Commission. (2009, October). Guides Concerning the Use of Endorsements and Testimonials in Advertising, 16 CFR Part 255.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Mind Mach 30*, 411–437.

General Data Protection Regulation. (2018, May). Article 25: Data protection by design and by default.

Goldstein, P. (2021, June 24). What is the state of government competency around AI? *FedTech Magazine*. https://fedtechmagazine.com/article/2021/06/what-state-government-competency-around-ai.

NextGov. (2020). Is the federal government ready for AI? https://www.govexec.com/insights/federal-government-ready-ai/.

Harari, Y. N. (2018). *21 Lessons for the 21st century*. Jonathan Cape.

Harris, T. (2016, May 18). How technology hijacks people's minds: From a magician and Google's design ethicist. *Medium Magazine*. https://medium.com/thrive-global/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors 57*(3), 407–434.

Kleinig, J. (2020). Loyalty. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)*. https://plato.stanford.edu/archives/win2020/entries/loyalty/.

Lam, M. S., Campagna, G., Xu, S., Fischer, M., & Moradshahi, M. (2019). Protecting privacy and open competition with almond. *XRDS: Crossroads, The ACM Magazine for Students 26*, 40–44.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck Lepri, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology 31*, 611–627.

List, C. (2013). Social choice theory. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2013 Edition)*. https://plato.stanford.edu/archives/win2013/entries/social-choice/.

National Security Commission on AI (National Security Commission on Artificial Intelligence). (2021, April 22). Final Report. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

National Cyber Security Centre. (2019, May). Secure design principles. https://www.ncsc.gov.uk/collection/cyber-security-design-principles.

Niker, F., & Sullivan, L. S. (2018). Trusting relationships and the ethics of interpersonal action. *International Journal of Philosophy Studies 26*, 1–14.

Niker, F., Reiner, P. B., & Felsen, G. (2016). Pre-authorization: A novel decision-making heuristic that may promote autonomy. *The American Journal of Bioethics 16*, 27–29.

Niker, F., Reiner, P. B., & Felsen, G. (2018). Perceptions of undue influence shed light on the folk conception of autonomy. *Front Psychology 9*, 57–11.

Omohundro, S. M. (2008). The basic AI drives. *Frontiers in Artificial Intelligence and Applications 171*, 483–492.

Russell, S. (2019). *Human compatible: AI and the problem of control*. Viking Press.

Securities and Exchange Commission. (2019, July 12). Commission interpretation regarding standard of conduct for investment advisers. https://www.sec.gov/rules/interp/2019/ia-5248.pdf.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transp Res Part C Emerg Technologies 126*, 103069.

Stimmel Law. The fiduciary duty. (n.d.). https://www.stimmel-law.com/en/articles/fiduciary-duty.

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Vincent, J. (2020, January 7). White House encourages hands-off approach to AI regulation. *The Verge*. https://www.theverge.com/2020/1/7/21054653/america-us-ai-regulation-principles-federal-agencies-ostp-principles.

Veale, M., & Zuiderveen Borgesius, F. (2021, July 6). Demystifying the Draft EU Artificial Intelligence Act. https://doi.org/10.31235/osf.io/38p5f.

Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.

CHAPTER 17

INFORMATION MARKETS AND AI DEVELOPMENT

JACK CLARK

Technical Observatories for Better AI Governance

How we can use technical metrics to improve government assessment and oversight of AI

In 2010, cutting-edge computer vision systems could look at an image and come up with a set of five potential labels for it and be right 70 percent of the time. By 2021, computer vision systems can do the same task and are right 98 percent of the time. That is more than a 10X reduction in error rates in about a decade. Meanwhile, the costs of training computer vision systems have fallen dramatically. In 2017, training a system to a basic level of computer vision performance on a standardized task cost $1,100 to do on public cloud infrastructure; the same task now costs $7.43. Taken together, these metrics give us a sense for how the capabilities of computer vision have progressed, as well as how the costs of deploying them have fallen.

Yet these kinds of metrics, while used widely by technical practitioners, are rarely integrated into the decision-making processes of governments. Today, governments use certain metrics to help orient themselves to specific issues with specific technologies (e.g., using the disengagement rate of self-driving cars to understand progress there, or to conduct bias-centric analysis of facial recognition systems), but do not use metrics in a systematic, holistic way to orient them to progress and issues in the field as a whole. To have a better regulatory environment for AI development we should seek to build infrastructure that can let governments move from periodic, ad-hoc assessments of AI systems to continuous, holistic analysis of the field as a whole.

In this chapter, I will discuss why I was able to access the technical data discussed above and what I and others in the AI community use this information for, why governments are currently poorly placed to use data like this, and some actions that can be taken to improve

this situation. Ideally, this chapter will shed some light on the relationship between information about the technical progress of AI and regulatory bodies, and propose some ways to more tightly integrate the two to improve AI governance and AI development.

This chapter is structured into two parts:

1) How the AI community uses metrics and measures to understand its own progress, and
2) How governments could integrate these metrics and measures into policymaking.

The main points made in this chapter are that AI metrics (and their associated datasets) give policymakers an evidence base which they could more closely integrate into regulatory activities. Doing so will give us broader awareness of the AI sector, which can help create a better and more informed regulatory system. Specifically, by integrating metrics into policymaking, we claim:

- Metrics make AI development legible to a non-expert, third party observer (such as a policymaker).
- The development of metrics helps surface valuable information about progress in a given AI field, which can orient people as to the likelihood of imminent real-world impact.
- Metrics and datasets help to provide early identification of flaws in AI technologies, which can better equip policymakers to understand where to regulate (to prevent harms) and where to invest resources (to maximize benefits).
- Governments should move from periodic, one-off assessments of specific AI capabilities to continuous assessment of the development of AI metrics and datasets.

How the AI Community Uses Metrics and Measures to Understand AI Progress

I was able to derive the facts about computer vision error rates listed in the opening paragraph of this chapter very easily. This is because the computer vision community has been testing out increasingly advanced systems on the same, underlying dataset, called ImageNet. This means that there is highly reliable data available for modeling the progress of computer vision. By testing systems on ImageNet and then amalgamating the metrics yielded by these tests, researchers can analyze not only individual breakthroughs, but use ImageNet as a proxy measure for trends in the field as well.[1] This information is periodically consolidated by a range of different services and organizations, including the AI Index at Stanford University (which I co-chaired in 2021) and the website Papers with Code (see Figure 17.1).

These metrics are how the AI field orients itself—if researchers make a meaningful breakthrough on a hard benchmark like ImageNet, that in turn incentivizes other researchers to work on it. Similarly, if a benchmark proves challenging to make progress on, that can sometimes force the development of novel techniques to ultimately improve performance against it. For example, Microsoft Research developed "Residual Networks" and applied them to ImageNet in 2015 as a way to set a new state-of-the-art score on ImageNet; so-called

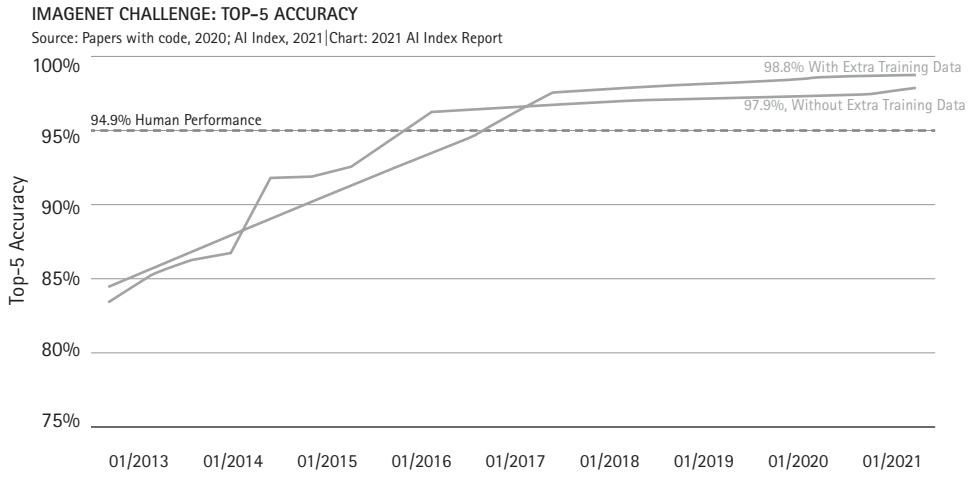

FIGURE 17.1 Imagenet Challenge: Top 5 Accuracy

Source: Papers with Code, 2020; AI Index, 2021 (Chart: 2021 AI Index Report)

ResNets were notable for stacking hundreds of layers in the network, where prior ones dealt in tens.[2] These metrics tend to exhibit a "rich get richer" phenomenon; that is, once a dataset starts to be used by a few researchers to test systems, other researchers will start to use that dataset and evaluation metric as a baseline for evaluating the progress of their own systems. AI, as a field, is driven forward by competition among researchers to set a new "state-of-the-art" on benchmarks like ImageNet.

The most useful aspect of this is the inherent openness of the benchmarks as well as the research papers that test systems out on them—the datasets circulate formally and informally among the research community with many being hosted on open-to-the-public sites like GitHub or Kaggle, while the benchmarks and research techniques applied to them are frequently documented in open access research papers published to repositories such as arXiv. Once published, they can be analyzed by other researchers, with performance on a given benchmark plotted over time giving us a sense of aggregate progress in the field.

Add it all up and you have a very interesting trait inherent to the AI community—many aspects of AI progress are inherently *legible to a third-party observer*. One simply needs to identify a problem (e.g., assigning the correct label to an image), then find a dataset that can test for this problem (e.g., ImageNet, which includes millions of images and their corresponding labels), and then search for papers that try to solve the problem using a common dataset and evaluation criteria.

How the AI community regulates and develops itself via measurements

When we study the AI research community, it is clear that metrics and their associated datasets are fundamental to not only technical progress in the field, but also the development

of metrics and measures to help the field understand its own blind spots and to further the development of ethical approaches to AI development. This cuts both ways—people can design metrics which would let them do things of arguable ethical legitimacy, and people can also design metrics to highlight ethical issues in systems. When we zoom out and look across the field, we see that progress is determined by performance of systems relative to technical benchmarks, and these various technical benchmarks and results become data used to analyze the ethical aspects of the field (both in terms of system performance and system failures).

In this section, I will highlight some measures, relate them to progress in the field, as well as the function they play in steering and regulating the development of the field.

Computer Vision: ImageNet and its variations

First introduced in 2009, ImageNet is a dataset containing more than one million labeled images split across 1,000 distinct categories. In 2010, ImageNet started to be used as the dataset for a competition that sought to evaluate how well computer vision systems could label and locate objects in photos. In 2012, a team from the University of Toronto used deep learning techniques (which were novel at the time) to score around 85 percent accuracy on classifying images—the second-place team got around 75 percent. This substantial performance jump led to many researchers paying more attention to the deep learning techniques used to set the impressive performance figure. (Note: when I refer to the ImageNet benchmark, I refer to scores set against the dataset used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC),[3] and when I refer to the ImageNet dataset, I'm referring to the underlying dataset which ILSVRC is a subset of).

Since then, ImageNet has been used as one of the main datasets and associated metrics that people use to understand the progress of the computer vision field as a whole. Other AI developers have also used ImageNet for strategic purposes. Microsoft, for example, chose to reveal its "residual network" technology by using the technology in a submission to the ImageNet competition in 2015—a competition which Microsoft won, drawing researchers' attention to the technical innovation it had made.

However, since then, AI researchers have encountered diminishing returns on using ImageNet to assess the development of computer vision, and scores have begun to asymptote as they approach the theoretical maximum of 100 percent. So, what have researchers done? They have created *harder variations* of ImageNet—datasets that are designed to test for specific known weaknesses in systems that score well on datasets like ImageNet.

These datasets, which are variants of the original ImageNet dataset, include:

- ImageNet-A: A dataset that, according to the authors, "is like the ImageNet test set, but it is far more challenging for existing models" (https://arxiv.org/abs/1907.07174).
- ImageNet-O: A dataset containing challenging out-of-distribution datasets.
- ImageNet-P: A dataset containing various transformations of ImageNet images, to help researchers understand the robustness of these systems (https://arxiv.org/abs/1903.12261).

Because these datasets are new, they have yet to generate much information. But in time, we can expect some researchers will test some systems against these datasets, as well as

against ImageNet itself. In this way, we can see how the creation of a dataset and a testing methodology can:

- provide a venue for researchers to understand the state-of-the-art for a given capability;
- serve as an early-warning system for the arrival of field-changing techniques (e.g., the use of deep learning in 2012 and the use of residual networks in 2015);[4] and
- encourage the development of subsequent, harder datasets based on the originating dataset, making it easier for researchers to test for the strengths and weaknesses of different techniques.

Natural language processing: Moving from discrete tests to evaluation suites

Although distinct, one-off evaluations of a particular AI system (or a particular technique applied to a specific dataset) can be useful, another trend has been driving development of measurement and assessment within the AI field: the shift from using single tests to evaluate systems to using suites of tests. This phenomenon is most visible in natural language processing, where in recent years the "GLUE" and "SuperGLUE" benchmark suites have been developed.

GLUE, first proposed in 2018, is a collection of distinct tests for evaluating the capabilities of Natural Language Processing (NLP) systems. GLUE lets developers test a single NLP system on a variety of distinct tasks, giving them both granular task-specific scores, as well as a blended score for their overall performance across these diverse task domains. In this way, GLUE can serve as a proxy for the improvement in capabilities on a broad set of dimensions, while also being intrinsically harder to game than a single metric that depends on a single evaluation approach.

In addition, because GLUE makes both sets of numbers available, it can be used to generate information about the broad capabilities of a system (via the single, blended number), as well as the distinct performance properties for specific tasks. Notably, GLUE was shortly superseded by SuperGLUE, which was first proposed in 2019. Like GLUE, SuperGLUE focuses on a set of harder tasks than those found in GLUE. SuperGLUE was also created because, per the authors: "Performance on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research" (Wang et al., 2020).

Where the AI community falls short, today, on measurements and evaluations

Because the AI community's use of metrics leads to continual generation of data, we can analyze the way researchers use metrics in the aggregate. By doing this, we can identify a couple of key changes the AI community is currently going through: first, that its benchmarks are becoming outmoded at an increasing rate and, second, that we are seeing the emergence of a new class of benchmarks oriented around measuring the social and normative aspects of AI work.

Our benchmarks are becoming irrelevant at an increasing rate

Due to the rapid maturation of the AI field—specifically, research into deep learning—the 2010–2020 period saw the invention of numerous systems that meaningfully improved the state-of-the-art on a range of metrics. But beyond just improving the state-of-the-art, we also started to see a reduction in the time it would take to saturate or outpace a benchmark. As an analysis from the *Dynabench* paper[5] shows, we have gone from a world where a benchmark (for instance, MNIST) will take decades to become saturated, to one where benchmarks can be saturated within a year or less (see Figure 17.2).

This poses a challenge and an opportunity to AI researchers and the people that would like to govern and understand then: how can we increase the rate at which we develop benchmarks to keep being able to generate signals about the progress in the field? And are there opportunities to explore where these benchmarks can be created with the intention of encouraging progress towards specific, narrow policy or regulatory goals?

The emergence of new AI ethics papers and benchmarks

In recent years, there has been an explosion in the number of papers being published that deal with the ethical aspects of AI technology (see Figure 17.3).

This rise in publishing of ethics papers has been coupled with an increase in the focus on—and availability of—datasets and associated metrics for evaluating societally- or policy-relevant aspects of AI systems.[6] This has included the emergence of datasets meant to help us study the fairness and bias aspects of AI systems, as well as more standard technical robustness (such as the earlier discussed ImageNet alternative metrics). There are also

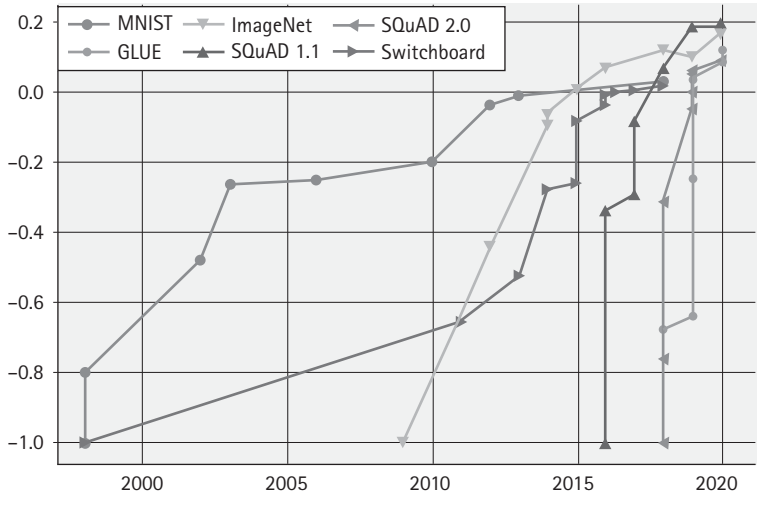

FIGURE 17.2 In recent years, it has taken less and less time for AI systems to exceed human performance on newly created benchmarks

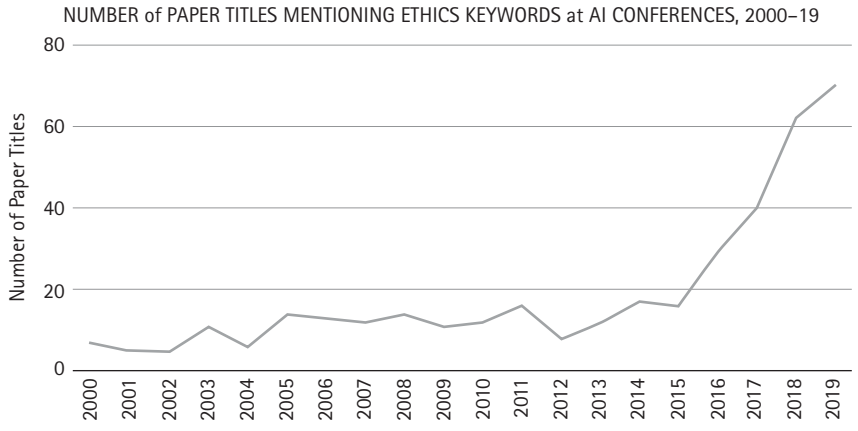

FIGURE 17.3 Number of Paper Titles Mentioning Ethics Keywords at AI Conferences, 2000–2019

Source: Prates et al., 2018; AI Index, 2021 (Chart: 2021 AI Index Report)

approaches being developed[7] to analyze the computational efficiency of AI systems, which generates data to let us think about the potential environmental effects of the field as it develops.

These datasets and metrics typically take two forms: either, they increase the amount of data available for a given sub-slice of the data distribution (e.g., Facebook's "Casual Conversations"[8] dataset seeks to create a demographically representative dataset of video conversations among Americans of a range of cultural backgrounds, racial identities, ages, and genders), or involves creating datasets that serve to evaluate a model for its classification *preferences* with regards to different datasets (e.g., StereoSet, which measures stereotypical bias in pretrained language models[9]).

Both types of datasets and metrics can generate information about some of the downstream impacts of AI systems, such as their likelihood of misclassifying people from certain backgrounds, or likelihood of having "blind spots" in their ability to perform well for some parts of the data distribution.

How Governments Could Integrate These Metrics and Measures into Policymaking

Given that the AI community is continually generating information about technical progress, as well as datasets and metrics that give an indication of where progress is slow or where contemporary AI systems are challenged, we might ask the question: how can we ensure this information makes its way to policymakers?

This is a valuable thing to consider: policymakers have the ability to set budgets, marshall resources for increasing (or reducing) funding for different streams of scientific research, and are tasked with thinking about national and economic security issues for their citizens.

Therefore, we should ask why, given all this information in the AI community, we don't see more of it directly inform policymaking.

Why governments are not (yet) able to regulate AI via measurements, and what to do about it

Today's policy infrastructures are (mostly) blind to advances like ImageNet, and to the many ways in which the AI community uses assessments and measurements to regulate itself. Although the AI research community spent the first half of the 2010s re-orienting itself around a new class of powerful, capable vision algorithms, governments were mostly unaware that such a significant change was taking place—or if they were aware, it happened at a significant time delay.

This is because most governments lack the bureaucratic equivalent of the sensors that would cue them to these changes. This had led to something deeply unfortunate occurring—policymakers are regularly *surprised* by the consequences of such significant technical progress. Such surprises can come about as a consequence of faster-than-anticipated technical progress leading to the rapid deployment of a new technological capability, or a technological accident or abuse leading to a surprising event which triggers reactive legislation. And as people who work in policy know, a surprised policymaker is more prone to carrying out hasty or poorly thought-out regulation than a calm one.

To get a sense of what we mean, we can review some cases where policymakers have been surprised by AI capabilities or actions in recent years.

Deepfakes

In recent years, advancing AI techniques have been used to create synthetic images, video, and audio that can impersonate real people. This has generated significant alarm among policymakers, who have worried about the impact entirely fake media could have on democracy.[10] These kinds of techniques have been in development since 2014, with the technical trajectory becoming clear by 2016 (see Figure 17.4), yet the first congressional hearing on the subject took place in 2019.[11]

If the technical community was generating useful technical information about deepfakes from 2014, with trends becoming clear by around mid-2016 (see Figure 20.4), we might ask the question of how we could better integrate these technical artifacts into policymaking.[12] If governments thought about deepfakes at an earlier point, it is likely they would have increased funding into researching the development and detection of them earlier.

Fairness

Modern machine learning systems can exhibit a range of fairness issues which have downstream impacts of deployment. Specifically, machine learning algorithms tend to soak up and exhibit the biases in the data they were trained on. Researchers and the general public have been studying these issues of fairness for some years. In 2015, a user of Google photos found that the application consistently miscategorized pictures of Black people as gorillas,[13] and in

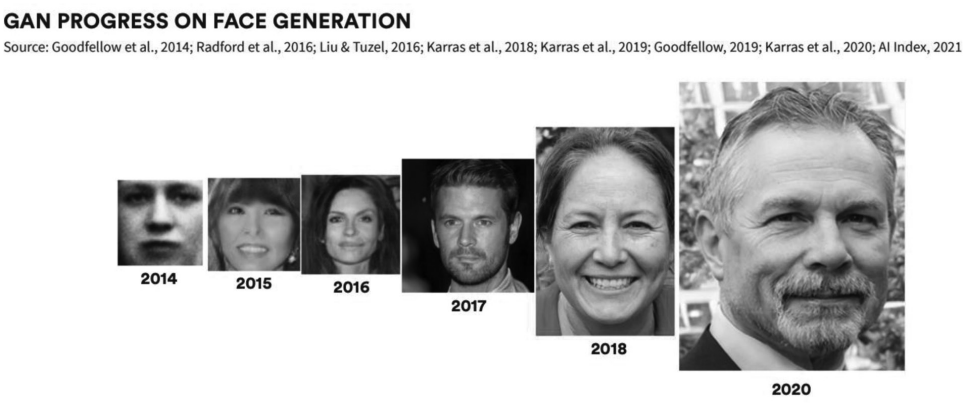
GAN PROGRESS ON FACE GENERATION
Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021
2014
2015
2016
2017
2018
2020

FIGURE 17.4 GAN Progress on Face Generation

Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021

2016, researchers demonstrated that text-analysis systems developed gender-specific biases. These concerns were laid out in a paper whose title is somewhat self-explanatory: "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings."[14]

Since then, there have been a wide range of datasets and metrics that researchers have used to highlight the biased and unfair ways in which systems classify things and generate synthetic outputs that relate to these things. At the same time, there are no standards that have emerged in the community for how to test for issues like fairness in a standardized way. All the while, policymakers are increasingly demanding that we assess AI systems for traits relating to concerns like fairness, security, reliability, and so on. This creates a gap between policymakers and AI developers, where a lack of calibration about the technical state of the art among policymakers has led to them having miscalibrated expectations about what is possible today, leading to the design of a less effective regulatory regime.

One example of this potential miscalibration is in relation to the fairness and bias in AI systems. An April 2021 draft of European Commission legislation for the regulation of AI notes that "high-risk AI systems" will need to have their datasets come under data governance and management practices which shall include the "examination in view of possible biases" of system datasets, as well as a desire by policymakers for these datasets to be "relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used."[15] This desire from policymakers doesn't quite match up to the maturity of the current AI field with regard to being able to analyze systems for bias—essentially, policymakers are writing legislation that assumes researchers have access to technical evaluation capabilities which don't currently yet exist. While this legislation could serve to motivate the development of such techniques, its position relative to the state of the field is still notable—policymakers desire approaches regarding biases that aren't (yet) technically possible, suggesting a mismatch of priorities between policymakers and developers. How might the legislation have been written if the European Commission had access to in-house technical teams which could give it a better understanding of how fairness evaluations are carried out today?

So, how might governments better integrate the information that exists around AI systems into decision making, or how might they stimulate the creation of information where none exists? To understand this, we should first briefly outline how governments deal with technical information relevant to AI today, then discuss some of the ways we could create new systems in government to handle this information.

How governments approach questions of measurement and assessment today

Today, governments typically approach questions of measurement and assessment for a given technology through the following methods:

- **Expert convening:** They will convene experts to synthesize information and produce a report including recommendations. Recent examples of such convenings include the National Security Commission on AI (NSCAI)[16] in the United States, whose recommendations are currently influencing government strategy. An earlier example might be the *Lighthill Report*,[17] a colloquial name given to a report compiled for the UK government in the 1970s on AI advances by James Lighthill.
- **Capability-based analysis:** Governments will periodically identify specific technical capabilities as being important to the broader evaluation of peer competition, and then will use these technical capabilities as proxies for broader competition across nations. Some examples here would include the Space Race in the 1950s to 1980s, when America and Russia competed to field specific technical capabilities (e.g., putting a person on the Moon). Other examples include nations racing against each other to sequence the Human Genome and, more recently, the United States, China, and Russia competing according to how adept each party is at developing the technology for loosely coordinated drone swarms. Competition can look very different depending on what you are measuring; for instance, by some metrics the United States is producing very strong publications in Natural Language Processing, while China is producing stronger publications in Computer Vision and AI as a whole.[18]
- **Regulatory compliance-based measurement analysis:** For more mature technologies, governments also assess how well they conform to regulation by measuring and assessing them. For instance, food standards are a type of metric which governments will test foods against, allowing them to use a measure (e.g., sodium content) to let them regulate towards a national policy goal (e.g., a healthier population).

One way governments could integrate measures from the AI community for better governance

In this chapter, we have surveyed how measures and assessments help the AI community orient itself around technical progress and also surface governance challenges for itself. At the same time, we have discussed how governments typically try to surface such measures themselves, either through convening experts (to synthesize metrics for them), or to pick

a highly specific metric tied to a capability (e.g., the ability to coordinate a fleet of drones). Governments only tend to use measures in a consistent and continuous way once the underlying industry has matured and been fully regulated—which is not yet the case for AI.

So, what are some potential solutions? There are a couple of approaches that can be adopted:

- Increase the resources of the regulators so that they are able to create more measures, which will help generate information.
- Increase the amount of measurement and assessment work being done broadly, then coordinate the work with demands from regulators and other parties.

For the first, increasing the resources of regulators so that they are able to create more measures, one approach is outlined in *Regulatory Markets for AI Safety*. The idea is that we can harness the measures and metrics of the AI research community to create a new regulatory layer, whereby the government would create a private market for regulators, which would then compete with one another to offer the fullest possible suite of regulatory tests for their customers (firms doing business in the market).

In a *Regulatory Markets* structure, the government retains the right to grant or revoke licenses for firms operating as private regulators, and the firms would have an incentive to compete with one another to design rapid and adaptable regulatory schemes that their customers could use. A regulatory markets approach could create an incentive for firms to create more ways to assess issues of fairness and representation in AI algorithms, as firms might want to be able to show a government they have used the most comprehensive evaluation methodology on offer in the market.

However, a regulatory markets approach is likely necessary but not sufficient for having a well-regulated AI ecosystem. This is because while it harnesses the AI community's traits to generate more information, it doesn't create as much direct regulatory capacity or awareness within governments themselves, which could represent a weakness—specifically, via increasing the likelihood of "regulatory capture" of either the regulatory bodies, or the government agencies that regulate the regulators.

Increase the amount of measurement and assessment work being done broadly, then coordinate the work with demands from regulators and other parties

Another approach might be to fund the creation within government of new capabilities for assessing and measuring the technical development of AI systems, then aligning these measures with systems of governance—such as, relating these metrics to funding bodies, tying these metrics into critical national economic or security functions, and providing a resource within government for understanding AI advancement. Although this approach sounds intuitive—after all, why wouldn't we want governments to carefully assess and measure AI technology?—it is actually not very common. Some parts of some governments are built to deal with technical metrics and measures—for example, the National Institute of Standards and Technologies (NIST) in the United States periodically conducts technical assessments and measurements of AI systems, such as the Face Recognition Vendor Test (FRVT).

However, no single government organization exists that is continuously measuring and assessing technical progress within artificial intelligence. This means that when agencies do analyze AI, they tend to do so through the lens of a specific, niche issue (such as NIST's analysis of facial recognition), rather than measuring things as part of a holistic enterprise. (Note: Some exceptions exist, such as the mothballed Office of Technology Assessment in the United States.[19])

Instead, we might propose that governments fund an agency, or a part of an existing agency, to do continuous analysis of the technical metrics and measures being developed by artificial intelligence researchers. This is not as intimidating an undertaking as it sounds because AI naturally produces measures and metrics, and so it is possible to extract granular metrics from much of AI research. Additionally, we can also infer the *absence* of metrics. For instance, governments may want to ensure there are more tests and metrics created to assess things like the bias of a given AI system, along with raw performance; by continuously analyzing the research literature, it would be easy to spot gaps in metrics. As we have discussed in this chapter, areas as diverse as computer vision, fairness-based analysis of systems, and the advance of synthetic media can all be analyzed within a metrics-based approach. By having the government more closely monitor and synthesize these technical metrics, we would ensure parts of government were more aware of problems (such as the bias-related issues shown by various fairness measurements), were more sensitive to periods of significant technical change (such as the performance improvement on ImageNet in recent years), and could understand areas where progress is lagging or non-existent. This could yield a range of beneficial outcomes, ranging from increased awareness of potential weaknesses in the technology, to awareness of when parts of it were rapidly maturing. It is likely that by generating this information in the public domain, other ancillary benefits could crop up as well.

Conclusion

In this essay, we have discussed some of the technical metrics and measures by which the AI community orients itself around matters of technical progress and challenges. We have also discussed how these measures can sometimes overlap with governance, either by creating information to enable better governance, or by highlighting issues that require governance. Future questions might include: what is the precise bureaucratic means by which we could implement better technical measurement and assessment within a government? Another question would be how to identify the initial things such an organization could measure, and what the main goal should be?

Notes

1. Of course, some individual metrics or measures can be misleading—the "progress" portrayed in the metrics may not correlate to the real world. However, if enough metrics are synthesized and viewed in the aggregate, it is possible to see broader trends in the research, avoiding dependency on any single benchmark.

2. He, K., et al. (2015). Deep residual learning for image recognition. *arXiv*, December. https://arxiv.org/abs/1512.03385.
3. For more information about the ImageNet competition, refer to https://image-net.org/challenges/LSVRC/.
4. For more discussion of this idea, see Cremer, C. Z., & Whittlestone, J. (2021). Artificial canaries: Early warning signs for anticipatory and democratic governance of AI. International *Journal of Interactive Multimedia and Artificial Intelligence 6*(5), 100–109. https://www.ijimai.org/journal/bibcite/reference/2905.
5. Kiela, D. (2021). Dynabench: Rethinking benchmarking in NLP. *arXiv*, April. https://arxiv.org/abs/2104.14337.
6. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-Chapter-5.pdf.
7. See Hernandez, D., & Brown, T. B. (2020). Measuring the algorithmic efficiency of neural networks. *arXiv*, https://arxiv.org/abs/2005.04305.
8. https://ai.facebook.com/datasets/casual-conversations-dataset/.
9. Nadeem, M., et al. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv*, April, https://arxiv.org/abs/2004.09456.
10. This could be an example of how policymakers themselves are, even when aware of technical trends, miscalibrated as to where the harms are occurring; while politicians worried in 2019 about the use of deepfakes for disinformation, a report by deepfake-detection company Sensity found that 96 percent of deepfakes in circulation were pornographic in nature. Sensity. (2019). The state of deepfakes: Landscape, threats, and impact. *Sensity.ai*. https://sensity.ai/reports/.
11. https://www.congress.gov/event/116th-congress/house-event/109620.
12. From *AI Index 2021*.
13. BBC News. (2015). Google apologises for Photos app's racist blunder. *BBC.com*, July 1. https://www.bbc.com/news/technology-33347866.
14. Bolukbasi, T., et al. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *arXiv*, July, https://arxiv.org/abs/1607.06520.
15. Proposal for a Regulation laying down harmonized rules on artificial intelligence, European Commision, Article 10—Data and data governance. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
16. https://www.nscai.gov/.
17. The *Lighthill Report* can be found at http://www.aiai.ed.ac.uk/events/lighthill1973/lighthill.pdf.
18. AI Definitions Affect Policymaking, Center for Security and Emerging Technology (CSET), 2020. https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf.
19. West, D. (2021). It is time to restore the US Office of Technology Assessment. *Brookings Institution*, February 10, https://www.brookings.edu/research/it-is-time-to-restore-the-us-office-of-technology-assessment/.

CHAPTER 18

ALIGNING AI REGULATION TO SOCIOTECHNICAL CHANGE

MATTHIJS M. MAAS

Introduction

How do we regulate a changing technology, with changing uses, in a changing world? As artificial intelligence ("AI") is anticipated to drive extensive societal change (Feldstein, 2023; Weidinger et al. 2023), the question of how we can and should reconfigure our regulatory ecosystems for AI-driven change matters today. In just the past decade, advances in both AI research and in the broader data infrastructure have begun to spur extensive take-up of this technology across society (Clark & Perrault, 2023; D. Zhang et al., 2021). As a "general-purpose technology" (Trajtenberg, 2018), AI's impact on the world may be both unusually broad and deep. Some suggest that over time it may even prove as "transformative" as the industrial revolution (Gruetzemacher & Whittlestone, 2022). This may provide grounds for anticipation—but also for caution. Most of all, it is grounds for reflection on the choices that societies want to make and instill in the trajectory of this technology, and their relation to it.

While still at an early stage of development, uses of AI technology are already creating diverse policy challenges. Internationally, AI is the subject to intense contestation. Further AI progress, along with the technology's global dissemination, are both likely to further raise the stakes. It is clearly urgent to reflect on the purposes and suitability of the regulatory ecosystem for AI governance. The urgent question is: when we craft AI regulation, how should we do so? Many AI governance approaches at both the national and international level remain hampered by siloed policy responses, tailored to individual AI applications (that is, they are technology-centric), or to AI applications' effects on individual legal fields or doctrines (that is, they are law-centric).

In contrast, this chapter argues that to craft adequate AI policies, a better regulatory perspective takes a step back, and first asks the questions: when we craft AI regulation, what are we seeking to regulate? And what are the ingredients for a more systematic regulatory template for crafting AI governance?

This chapter argues that to foster an effective AI regulatory ecosystem, policy institutions and actors must be equipped to craft AI policies in alignment with systematic assessments of how, when, and why AI applications enable broader forms of sociotechnical change (Maas, 2020). It argues that this sociotechnical approach complements existing technology-centric and law-centric models of AI policies; and that, supported by adequate institutional processes, it can provide an informed and actionable perspectives on when and why AI applications actually create a *rationale* for regulation—and how they are consequently best approached as *targets* for regulatory intervention. This enables more tailored policy formulation for AI issues, facilitates oversight and review of these policies, and helps address structural problems of accountability, (mis)alignment, and (lack of) information in the emerging AI governance regulatory ecosystem.

The chapter is structured as follows. It first (1) sketches the general value of a "change-centric" approach to AI governance. It then (2) proposes and articulates a framework focused on "sociotechnical change," and explores how this model allows an improved consideration of (3) when a given AI application creates a regulatory *rationale*, and (4) how it is subsequently best approached as a regulatory *target*, in relation to six distinct "problem logics" that appear in AI issues across domains. Finally (5), the chapter reflects on concrete institutional and regulatory actions that can be derived from this approach to improve the regulatory *triage*, *tailoring*, *timing and responsiveness*, and *design* of AI policies.

Towards Change-centric Approaches in an AI Regulatory Ecosystem

In response to AI's emerging challenges, scholars and policymakers have appealed to a wide spectrum of regulatory tools to govern AI. Unsurprisingly, recent years have seen increasing public demands for AI regulation (B. Zhang & Dafoe, 2020), and diverse new national regulatory initiatives (Cussins, 2020; Law Library of Congress, 2019).

Much academic work to date has focused on the regulation of AI within particular domestic regulatory contexts. For instance, this has explored the relative institutional competencies of legislatures, regulatory agencies, or courts at regulating AI (Guihot et al, 2017; Scherer, 2016). Others have emphasized the governance roles of various actors in the AI landscape (Leung, 2019), exploring for instance how tech companies' ethics advisory committees (Newman, 2020, pp. 12–29), AI company employee activists and "epistemic communities" (Belfield, 2020; Maas, 2019a), AI research community instruments (such as scientific conference research impact assessment mechanisms) (Prunkl et al., 2021), or private regulatory markets architectures (Clark & Hadfield, 2019), could all help shape AI regulation.

There is also a growing recognition of the importance of global coordination or cooperation for AI governance (Feijóo et al., 2020; Kemp et al., 2019; Turner, 2018). At the global level, most focus to date has been on the burgeoning constellation of AI ethics principles, which has sprung up over the last half-decade (Fjeld et al., 2019; Jobin et al., 2019; Schiff et al., 2020; Stahl et al., 2021; Stix, 2021). However, this is being increasingly complemented by proposals for harder law and regulation. Some approaches to global AI governance invoke existing

norms, treaty regimes, or institutions in public international law (Kunz & ÓhÉigeartaigh, 2021; Burri, 2017; Smith, 2020). Others have proposed entirely new international organizations to coordinate national regulatory approaches (Erdelyi & Goldsmith, 2018; Kemp et al., 2019; Turner, 2018, chap. 6), and have compared such centralized institutions for AI to more decentralized or fragmented alternatives (Cihon et al., 2020b, 2020a). Yet others have focused more on the role of soft law instruments (Gutierrez & Marchant, 2021; Gutierrez et al., 2020), international standard-setting bodies (Cihon, 2019; Lorenz, 2020), or certification schemes (Cihon et al., 2021), as well as proposing the adaptation of existing informal governance institutions, such as the G20 (Jelinek et al., 2020).

This is clearly a diverse constellation of efforts. Yet there are underlying classes and patterns in the prevailing approaches to AI regulation. For instance, one group of "technology-centric" approaches focuses on "AI" as an overarching class that should be understood—and regulated—holistically (Turner, 2018). While this is more reflective of the cross-domain application and impact of many AI techniques, however, this approach is not without problems. Critically, it is undercut by intractable debates over how to define "AI" (Russell & Norvig, 2016)—and by the fact that AI is not a single thing (Schuett, 2019; Stone et al., 2016).

This shortfall is addressed somewhat within the second type of "technology-centric" approach. Such an "application-centric"—or, as some call it, "use-case centric" (Schuett, 2019, p. 5)—approach seeks to unpack the umbrella term "AI," and split out the specific AI applications that regulation should focus on. In the past decade, many policy responses to AI have been sparked by one or another use case of AI technology—involving concrete issues that have been thrown up, or visceral problems that are anticipated—such as autonomous cars, drones, facial recognition, or social robots (Turner, 2018, pp. 218–219). As Petit puts it, this technology-centric approach to AI involves charting "legal issues from the bottom-up standpoint of each class of technological application" (Petit, 2017).

These application-centric approaches remain the default response to AI governance. However, like the AI-centric approach, this regulatory orientation also has shortfalls. For one, it emphasizes visceral or upsetting incidents or even imagined scenarios, and is therefore easily lured into regulating relatively rare edge-case challenges or misuses of the technology (e.g., the use of deepfakes for political propaganda) at a cost of addressing far more common but less visible use cases (e.g., the use of deepfakes for gendered harassment) (Liu et al., 2020), potentially exacerbating inequalities in regulatory attention. Moreover, the resulting AI policies and laws are frequently formulated in a piecemeal and ad-hoc fashion, which means that this perspective promotes siloed regulatory responses (Turner, 2018, pp. 218–219). A focus on individual applications also may inadvertently foregrounds technology-specific regulations even where these are not the most effective (Bennett Moses, 2013). It moreover induces a "problem-solving" orientation (Liu & Maas, 2021), aimed at narrowly addressing local problems caused (or envisioned) by the specific use case that prompted the regulatory process, rather than understanding commonalities in challenges (Crootof & Ard, 2021).

A distinct set of regulatory responses to AI are instead law-centric. They invoke what Nicolas Petit has called a "legalistic" approach, which "consists in starting from the legal system, and proceed[s] by drawing lists of legal fields or issues affected by AIs and robots" (Petit, 2017, p. 2). This approach segments regulatory responses by departing from AI's impacts on specific conventional legal codes or subjects (e.g., privacy law, contract law, the

law of armed conflict), or by exploring the ways in which these uses create questions about the scope, intersection, assumptions, or adequacies of existing bodies of law (Crootof & Ard, 2021).

To be clear, both application- and law-centric approaches to AI regulation have important insights and must play a role in any AI governance ecosystem. Nonetheless, they have their drawbacks. Importantly, AI governance proposals could be grounded in a better understanding of how AI applications translate into cross-domain changes, and how future capabilities or developments might further shift this problem portfolio (Maas, 2019b). An alternative is therefore to shift (or complement) these approaches with a framework that is not solely anchored in "new technology" (whether understood through the umbrella term of "AI," or with emphasis on numerous specific AI applications), nor on isolated legal domains—but which rather examines types of change. What impacts are we concerned about?

AI regulatory ecosystems require a protocol for considering *where, and why AI change warrants regulatory intervention, and how and when this regulatory intervention should take place*. It should be able to adequately identify when applications of AI create regulatory rationales, as well as the best levers to subsequently approach it as a regulatory target.

Can we reformulate an impact-focused approach for AI regulation that provides superior levers for regulation? To achieve this, this chapter draws on existing theories from the emerging paradigm of "techlaw"—"the study of how law and technology foster, restrict, and otherwise shape each other's evolution" (Crootof & Ard, 2021, n. 1; Ard & Crootof, 2020). In particular, it proposes to approach AI governance through the lens of "sociotechnical change" (Bennett Moses, 2007a, 2017). As such, this chapter will now turn to how this approach can bridge the gaps between technology/application-centric and law-centric approaches, by guiding reflection on when and why new AI applications require new regulation—and how the resulting regulatory interventions are best designed.

Reframing Regulation: AI and Sociotechnical Change

It should be no surprise that changes in technology have given rise to extensive scholarship on the relation between law and these new innovations. In some cases, such work has focused on identifying the assumed "exceptional" nature or features of a given technology (Calo, 2015). However, other scholars have influentially argued that it is less the "newness" of a technology that brings about regulatory problems, but rather the ways it enables particular changes in societal practices, behavior, or relations (Balkin, 2015; Bennett Moses, 2007b; Friedman, 2001).

That is, in what ways do changes in given technologies translate to new ways of carrying out old conduct, or to entirely new forms of conduct, entities, or new ways of being or connecting to others? When does this create problems for societies or their legal systems? Various scholars of law, regulation, and technology have emphasized the importance of focusing on the social or societal aspect of a new technology, rather than merely its presumed novelty. For instance, Brownsword, Scotford, and Yeung have highlighted three dimensions

of technological "disruption:" "legal disruption, regulatory disruption, and the challenge of constructing regulatory environments that are fit for purpose in light of technological disruption" (Brownsword et al., 2017, p. 7).

Accordingly, in her "theory of law and technological change" (Bennett Moses, 2007b), Bennett Moses has argued that questions of "law and technology" are rarely, if ever, directly about technological progress itself (whether incremental or far-reaching) (Bennett Moses, 2007b, p. 591). To the contrary, she argues that, from a regulatory perspective, "[t]echnology is rarely the only 'thing' that is regulated and the presence of technology or even new technology alone does not justify a call for new regulation" (Bennett Moses, 2017, p. 575). Instead, she argues that lawyers and legal scholars who examine the regulation of technology are focused on questions of "how the law ought to relate to activities, entities, and relationships made possible by a new technology" (Bennett Moses, 2007b, p. 591).

In doing so, she calls for a shift in approach from "regulating technology" to "adjusting law and regulation for sociotechnical change" (Bennett Moses, 2017, p. 574). This shifts the focus to patterns of "socio-technical change" (Bennett Moses, 2007b, pp. 591–592, 2007a): instances where changes in certain technologies actually expand human capabilities in ways that give rise to new activities or forms of conduct, or new ways of being or of connecting to others (Bennett Moses, 2007b, pp. 591–592).

As such, the question of governing new technologies is articulated not with reference to a list of (sufficiently "new") technologies (Bennett Moses, 2017, p. 576), but is relatively "technology-neutral." It is this functional understanding of "socio-technological change" that informs more fruitful analysis of when and why we require regulation for new technological or scientific progress. Most importantly, this lens can underlie a more systematic examination of which developments in AI technology are relevant for a regulatory system to focus on and prioritize.

AI as Regulatory Rationale

However, what types of sociotechnical changes (e.g., new possible behaviors or states of being) give rise to regulatory rationales? When a technology might create an opportunity for certain problematic behavior, does that opportunity need to be acted upon, or can the mere possibility of that behavior constitute a regulatory rationale? Can sociotechnical changes be reliably predicted or anticipated? This requires a more granular understanding of the dynamics of sociotechnical change.

Varieties of sociotechnical change

When and why do AI capabilities rise to a problem that warrants legal or regulatory solutions? It is important to recognize that not all new scientific breakthroughs, new technological capabilities, or even new use cases will necessarily produce the sort of "sociotechnical change" that requires regulatory responses.

In practical terms, this relates to the observations that new social opportunities or challenges (and therefore governance needs) are not created by the mere fact of a

technology being conceived, or even prototyped. The threshold rather comes when they are translated into new "affordances" for some actors. Affordances are relationships "between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used" (Norman, 2013, p. 11). AI affordances can be new types of behavior, entities, or relationships that were not previously possible (or easy), and which are now available to various actors (Liu et al., 2020).

How does technological change translate into sociotechnical change? When would this be disruptive to law? There are various affordances that new AI applications can create or enable, which could result in distinct types of relevant sociotechnical change (Maas, 2019c, p. 33; see also Crootof & Ard, 2021).

(1) *Allowing older types of behaviors to be carried out with new items or entities*, where AI progress provides new products or artifacts
   i. which are simply not adequately captured under existing technology-specific regulatory codes, or
   ii. which blur the boundaries between existing domains or regimes (e.g. a general-purpose language model seeing application in diverse industries; an autonomous weapon system seeing deployment in both civilian law enforcement and military), potentially causing problematic gaps, overlaps, or contradictions in how these behaviors are covered by regulation.

(2) *Absolute* or *categorical capability changes*, where AI progress expands the action space and "unlocks" new capabilities or behavior which were previously simply out of reach for anyone (e.g. highly accurate protein structure prediction), and which could be of regulatory concern for one of various reasons.

(3) *Relative capability changes*, where AI increases the prominence of a previously rare behavior, for instance because progress *lowers thresholds* or use preconditions for a certain capability (e.g., advanced video editing; online disinformation campaigns; algorithmic cryptographic tools), which was previously reserved to a narrow set of actors; or because progress allows the *scaling up* of certain existing behaviors (e.g., phishing emails).

(4) *Positional changes* amongst actors of a given type, (e.g. AI applications which shift power amongst particular states at the international level), while leaving the general "rules" or structure of that system more or less unaltered.

(5) *Changing structural dynamics* in a (international) society, for instance, by
   i. Shifting prevalent influence between *types* of actors (e.g. away from states and towards non-state actors or private companies);
   ii. Shifting the *means by* which certain actors seek to exercise "influence" (e.g., from "hard" military force to computational propaganda, or from multilateralism to "lawfare," as a result of new communications technologies increasing the scope, velocity and effectiveness of such "lawfare" efforts) (Dunlap, 2008, pp. 146–148); or
   iii. Altering the *norms or identities* of actors, and thereby changing the *terms* by which they conceive of their goals and orient their behavior.

As mentioned, there may be certain AI innovations or breakthroughs that do not create very large sociotechnical changes of these forms, even if from a pure scientific or engineering standpoint they involve considerable alterations to the state of the art.

Conversely, technological change or improvements also need not be qualitatively novel, dramatic, sudden, or cutting-edge for them to be sufficient to drive intense and meaningful change in balances of power, in societal structures (Cummings et al., 2018, p. iv), or one of the other affordances discussed. The question is therefore not only how large these sociotechnical changes are, but how, or whether, they touch on the general rationales for new regulatory intervention.

Mapping regulatory rationales

There are various accounts for when and why regulatory intervention is warranted by the introduction of new technologies. For instance, Beyleveld and Brownsword have argued that emerging technologies generally give rise to two kinds of concerns: "one is that the application of a particular technology might present risks to human health and safety, or to the environment [. . .] and the other is that the technology might be applied in ways that are harmful to moral interests" (Beyleveld & Brownsword, 2012, p. 35).

However, while these may be the most prominent rationales, the full scope of reasons for regulation may extend further. In a non-technology context, Prosser has argued that regulation, in general, has four grounds: "(1) regulation for economic efficiency and market choice, (2) regulation to protect rights, (3) regulation for social solidarity, and (4) regulation as deliberation" (Prosser, 2010, p. 18).

How do these regulatory rationales relate to technological change? As Bennett Moses (2017, p. 578) notes, all four of these rationales can certainly become engaged by new technologies. That is, new technologies (or new applications) can:

(1) Create sites for *new market failures*, warranting regulatory interventions such as technical standards or certification, to ensure economic efficiency and market choice, and remedy information inadequacies for consumers;
(2) Generate many *new risks or harms*—either to human health or the environment, or to moral interests—which create a need for regulation to protect the rights of these parties (e.g., restrictions of new weapons; the ban on human cloning);
(3) Create concern about *social solidarity*, as seen in concerns over the "digital divide" at both a national and international level, creating a need for regulation to ensure adequate inclusion; and
(4) Create sites or pressures for the exertion of *proper democratic deliberation* over the design or development pathways of technologies (Bennett Moses, 2017, pp. 579–583).

To be sure, as a technology-centric approach would note, these cases all involve new technologies which require regulation. However, Bennett Moses argues that in each of these situations, it is not the involvement of "new technology" per se that provides any special rationale for regulation above and beyond the resulting social changes (e.g., potential market failures, risks to rights, threats to solidarity, or democratic deficits) that are at stake (Bennett Moses, 2017, p. 583). We are not worried about technology; we are worried about its *effects* on ourselves.

As such, the primary regulatory concern is over the emergence of the "sociotechnical" effects that occur. This conceptual shift can help address one limit that regulatory or

governance strategies encounter if they focus too much or too narrowly on technology. As Bennett Moses argues:

> treating technology as the object of regulation can lead to undesirable technology specificity in the formulation of rules or regulatory regimes. If regulators ask how to regulate a specific technology, the result will be a regulatory regime targeting that particular technology. This can be inefficient because of the focus on a subset of a broader problem and the tendency towards obsolescence.
>
> (Bennett Moses, 2017, p. 584)

As such, a sociotechnical (rather than a technology-centric) approach helps keep into explicit focus the specific *rationales* for governance in each use case of AI: on what grounds and when regulation is needed and justified?

These four accounts of rationales are valuable as a starting point for AI regulation. However, we can refine this account. For one, it is analytically valuable to draw a more granular distinction between (physical) harms to human health or the environment, and (moral) harms to moral interests (Beyleveld & Brownsword, 2012).

Moreover, these categories all concern rationales for governance to step in, in response to sociotechnical changes that are affecting society (i.e., the regulatees) directly. However, there may also be cases where AI-enabled sociotechnical change creates an indirect regulatory rationale because it presents some risk directly to the existing legal order charged with mitigating the direct risks. In such cases of "legal disruption" (Liu et al., 2020; Maas, 2019c), sociotechnical change can produce a threat to the regulatory ecosystem itself. This can be because these tools allow intended regulatees to more effectively challenge or bypass existing laws, resulting in potential "legal destruction" (Maas, 2019c). Alternatively, it can result because certain AI tools can drive "legal displacement" (Maas, 2019c), by offering substitutes or complements to existing legal instruments, in governmental efforts in shaping or managing the behavior of citizens (Brownsword, 2019).

Drawing together the above accounts, one might then speak of a regulatory rationale for an AI system or application, whenever it drives sociotechnical changes (new ways of carrying out old behavior, or new behaviors, relations, or entities) which result in one or more of the following situations:

(1) New possible market failures;
(2) New risks to human health or safety, or to the environment;
(3) New risks to moral interests, rights, or values;
(4) New threats to social solidarity;
(5) New threats to democratic process; or
(6) New threats to the coherence, efficacy or integrity of the existing regulatory ecosystem charged with mitigating the prior direct risks (1–5).

All this is not to say that these rationales apply in the same way in all specific contexts. Indeed, they will be weighted differently across distinct legal systems and jurisdictions—as between domestic and international law. Nonetheless, they provide a rubric for understanding when or why we (should) want to regulate any new AI application—and a reminder that it is the sociotechnical changes, not the appearance of new technology in itself, that we are concerned about.

AI as Regulatory Target

Along with providing a greater grounding for understanding whether, when, and why to regulate new AI applications, a consideration of sociotechnical change can also shed light on the regulatory "texture" of the underlying AI capabilities—that is, its constitution as a "regulatory target" (Buiten, 2019, pp. 46–48).

That is, once regulators have established a regulatory rationale (i.e., they have asked "Do we need regulation? For what sociotechnical change? What regulatory rationale?"), they then face the question of how to craft regulatory actions in response. In considering AI applications as a *target* for regulation, a sociotechnical change-centric perspective must, interestingly, start by taking stock of the material aspects of a technology (as an artifact).

This is because material features certainly matter from the perspective of understanding key parameters for regulation, such as:

(1) *The technology's trajectory and distribution*: i.e., the state of leading AI capabilities (across its different sub-fields), possible and plausible rates and directions of progress given material constraints on the design space and process (Verbruggen, 2020); preconditions for acquisitions and use; and factors driving or inhibiting the technology's proliferation to various (types of) actors globally (Horowitz, 2018);
(2) *The technology's material "risk profile"*: how different AI paradigms or techniques—from Bayesian approaches to rule-based expert systems, and from evolutionary algorithms to reinforcement learning—can at times be associated with distinct types of ethics or safety issues (Hernandez-Orallo et al., 2020);
(3) *The political viability of regulation, given the technology's "regulation-tolerance/resistance" profile*: how certain features of the technology (e.g., the viscerality of weapons) affect stakeholder perceptions of the imminence of various applications, and of the need for urgent regulation (Crootof, 2019; Watts, 2015); and
(4) *Potential sites or vectors for regulatory leverage*: for instance, the degree to which proliferation of certain systems could be meaningfully halted through export control policies (Brundage et al., 2020; Fischer et al., 2021).

In these ways, material features certainly matter, especially when considering AI regulation at the global level. For instance, scholars have argued that the modern global digital economy, far from consisting solely of ethereal digital products ungraspable by law, is instead populated by distinct "regulatory objects," which vary in their degree of apparent "materiality" (from high-capital submarine cables and satellite launch facilities, to ethereal cloud services), and in their degree of centralization (from diverse suppliers of various "smart" appliances, to dominant social networks or computationally intensive search engine algorithms) (Beaumier et al., 2020). Critically, some of these may not easily be subjected to global regulation, but many can certainly be captured by various regulatory approaches and tools.

However, for regulatory purposes, a material analysis is not sufficient. A sociotechnical change-centric perspective on AI regulation rather can and should go beyond the technology itself and consider a broader set of "problem logics" in play.

We can fruitfully distinguish between: "ethical challenges," "security threats," "safety risks," "structural shifts," "common benefits," and "governance disruption." Differentiating amongst these ideal-types is valuable, as these clusters can introduce distinct problem logics, and foreground distinct regulatory logics or levers (see Table 18.1).[1]

It is important to note that this taxonomy is not meant to be mutually exclusive, nor exhaustive. It aims to capture certain regularities which help ask productive regulatory questions. For each category, we can ask: How does the AI capability produce sociotechnical change? Why does this create a governance *rationale*? How should this be approached as governance *target*? What are the barriers to regulation, and what regulatory tools are foregrounded?

There is insufficient space to go into exhaustive detail on each of these categories within the taxonomy. However, at glance, we can pick out a number of ways in which clustering AI's sociotechnical impacts by these problem logics can facilitate structural regulation-relevant insights for AI regulators.

These include consideration of various relevant aspects or regulatory parameters. For one, this model enables examination of the underlying *origins* of the sociotechnical challenge of concern, in terms of: (1) the *key actors* (e.g., principals, operators, malicious users) whose newly AI-enabled or -related behavior or decisions create the regulatory concern, and (2) those *actors' traits, interests, or motives* which drive the AI-sociotechnical-problem related behavior or decisions (e.g., actor apathy, malice, negligence, or the way the new capability sculpts choice architectures in ways shift structural incentives or strategic pressures).

This, in turn, can be linked to the *contributing factors* which sustain or exacerbate this sociotechnical impact, such as:

(1) The range and diversity of AI-related failure modes and issue groups (Hernandez-Orallo et al., 2020, p. 7), including emergent interactions with other actors (human or algorithmic) in their environment (Rahwan et al., 2019), as well as peculiar behavioral failure modes (Amodei et al., 2016; Krakovna et al., 2020; Kumar et al., 2019; Leike et al., 2017) (**safety risks**).
(2) Human over-trust and automation bias, which renders some AI systems susceptible to emergent and cascading "normal accidents" (Maas, 2018) (**safety risks**).
(3) The underlying "offense-defense balance" of AI scientific research (Shevlane & Dafoe, 2020), and how it evolves along with more sophisticated AI capabilities (Garfinkel & Dafoe, 2019) (**security threats**).
(4) The susceptibility of existing legal and regulatory systems themselves to "disruption" by AI uses, at the level of doctrinal *substance*, law-making and enforcement *processes*, or a legal system's *political* foundations and legitimacy (Liu et al., 2020; Maas, 2019c) (**governance disruption**).

Moreover, this model enables a study of the *barriers to regulation*; that is, the factors that determine the difficulty of formulating and implementing policy solutions, and which will themselves have to be mitigated or overcome if one seeks to achieve effective regulatory responses for AI. These include:

(1) Live societal debate or cross-cultural value pluralism (Gabriel, 2020), representing important disagreements over the values, interests, or rights at stake, and whether or

Table 18.1 Taxonomy of AI problem logics

Problem Logic and *questions*	Corresponding governance rationales	Examples in AI (*selected*)	Regulatory Surface (Origin; Contributing Factors; Barriers to Regulation)	Regulatory Approaches (*selected*)
Ethical challenges *What rights, values or interests does this threaten?*	• New risks to moral interests, rights or values • New threats to social solidarity • Threats to democratic process	• *Justice*: bias; explainability . . . • *Power*: facial recognition . . . • *Democracy*: AI propaganda . . . • *Freedom*: 'Code as Law'; 'algocracy' . . .	• O. Developer / user ***apathy*** (to certain implicated values) • BR. Underlying ***societal disagreement*** (culturally and over time) over how to weigh the values, interests or rights at stake	• *Product-focused:* Bans ('mend—or end'); 'machine ethics' • *Ex ante producer-focused:* oversight mechanisms; end-to-end auditing; ethics education for engineers; 'Value-Sensitive Design' processes • *Ex post principal*-focused: and accountability mechanisms
Security threats *How is this vulnerable to misuse or attack?*	• New risks to moral interests, rights or values • New risks to human health or safety	• *AI as tool*: Deepfakes; • *AI as attack surface*: adversarial input • *AI as shield*: fraudulent trading agents; UAV smuggling	• O. Attacker ***malice*** (various motives) • CF. Target ***apathy*** • CF. 'Offense-defense balance' of AI knowledge • BR. Target's intrinsic vulnerability (e.g., of human practices to automated social engineering attacks)	• *Perpetrator-focused*: change norms, prevent access; improve detection and forensics capabilities to ensure attribution and deterrence • *Target-focused*: reduce exposure; red-teaming; 'security mindset'
Safety risks *Can we rely on- and control this?*	• New risks to human health or safety	• *Unpredictability and opacity* • Environmental interactions • Automation bias and 'normal accidents' • 'Value misalignment'	• O. Actor ***negligence*** • CF. Behavioral features of AI systems (opacity; unpredictability; optimization failures; specification gaming) • CF. Human overtrust and automation bias • BR. 'Many hands' problem—long and discrete supply chains	• Relinquishment (of usage in extreme-risk domains) • 'Meaningful Human Control' (various forms) • Safety engineering (e.g. reliability; corrigibility; interpretability; limiting capability or deployment; formal verification) • Liability mechanisms and tort law; open development

how these should be weighted in the context of a specific contested AI application (**ethical challenges**);

(2) The disproportionately high costs of "patching" vulnerabilities of human social systems (e.g., our faith in the fidelity of human voices) against AI-enabled social engineering attacks, relative to past costs of patching "conventional" cybersecurity vulnerabilities by the dissemination of software fixes (Shevlane & Dafoe 2020, p. 177) (**security threats**); and

(3) The difficulty of foreseeing indirect effects of AI on the structure of different actors' choice architectures (van der Loeff et al., 2019; Zwetsloot & Dafoe, 2019)—and the difficulty of resolving those situations through any one actor's unilateral action (Dafoe, 2020), or to coordinate behavior in response (**structural shifts**).

Finally, on the basis of these factors, it allows a consideration of the types of *regulatory approaches* and levers that are highlighted and foregrounded for each of these challenges. It highlights the role of "mend-it-or-end-it" debates around algorithmic accountability (Pasquale, 2019), auditing frameworks (Raji et al., 2020), and underlying cross-cultural cooperation (ÓhÉigeartaigh et al., 2020) for AI's *ethical challenges*. It emphasizes perpetrator-focused and target-focused (e.g., "security mindset" [Severance, 2016]) interventions to shield against AI's *security threats*. Where it comes to programs aimed to guarantee *public goods* and common benefit of AI, it highlights the design of frameworks to ensure "AI for Good" interventions (Floridi et al., 2018; ITU, 2019), humanitarian uses (Roff, 2018, p. 25; but see Sapignoli, 2021), or redistributive guarantees such as a "Windfall clause" that sees tech companies pledge extreme future profits above a certain threshold towards redistribution (O'Keefe et al., 2020), amongst many others.

That is not to say that this framework provides conclusive recipes or roadmaps for regulation. Rather, it provides an initial structuring framework for thinking through common challenges across diverse regulatory domains charged with resolving questions around seemingly separate applications of AI (Crootof & Ard, 2021). Such an approach can at worst avoid duplication of effort, and at best can support the formulation and implementation of better, more effective, coherent and resilient policies for AI.

A sociotechnical-change-centric approach is not without its pitfalls or limits. Still, it can have various benefits in organizing and orienting an AI regulatory ecosystem. It prompts regulators to ask themselves: (1) when, why, and how a given AI application produces particular types of sociotechnical changes; (2) when and why these changes rise to create a rationale for governance; (3) how to approach the target of regulation. As such, this can be an important regulatory complement to the insights provided by—and the interventions grounded in—technology-centric or law-centric perspectives.

Implementation: AI Regulation through a Sociotechnical Lens

This lens of sociotechnical change does not provide single substantive answers for how to resolve each and every AI policy problem. However, it can help answer common recurring questions in AI policy around institutional choice and regulatory timing and design

(Bennett Moses, 2017, pp. 585–591). In particular, regulatory actors can use this approach to improve governance for AI challenges in terms of regulatory *triage*, *tailoring*, *timing and responsiveness*, and *design*.

Regulatory triage

For one, the sociotechnical-change-centric perspective on AI can help in carrying out regulatory *triage* (Hopster & Maas, 2023). This is not just of value to AI regulation: indeed, it fits in with a broader initiative in recent legal scholarship towards exploring questions of "legal prioritization" (Winter et al., 2021). Within the AI regulation ecosystem, this lens helps focus attention on the most societally disruptive impacts of the technology, and so helps re-focus scarce regulatory attention. This reduces the risk that regulatory attention is over-allocated on visceral applications of AI which may not ultimately prove scalable, or on philosophically or legally interesting "puzzles," at a cost of more opaque but more prevalent indirect impacts. Such triage can therefore be a valuable corrective to approaches that select, organize, or prioritize AI policy issues based on high-profile but non-representative incidents, popular-cultural resonance, or "fit" to pre-existing legal domains.

Moreover, if regulatory bodies focus less on the "newness" of AI technology as a class, or on the novelty of each and every latest AI application, and rather pay attention to which of AI's downstream sociotechnical impacts in fact create particular regulatory rationales, this will allow them to step back from a reactive firefighting mode, and help address, defuse or dissolve the so-called "pacing problem" (Marchant, 2011). Regulatory triage is also aided by the ways in which this framework can *expand regulatory actors' scope of analysis* in terms of which sociotechnical impacts qualify for regulatory consideration. While technology-centric approaches can highlight the direct challenges of AI (in the areas of *ethics*, *security*, and *safety*), the sociotechnical-change-centric perspective also allows regulators to consider interventions for various indirect sociotechnical changes, including the ways AI technology can *shift incentive structures*, could *offer beneficial opportunities* (that may need proactive policy to be fully realized as public goods), or *disrupt* the regulatory tools or systems which these regulators would rely on.

What does that entail in practice? Improving triage around AI regulation could involve (1) improving information infrastructures or "technical observatories" (Clark, 2024); Whittlestone & Clark, 2021), to not only equip regulators with relevant and up-to-date technical information around AI techniques and applications, but also think through how these feed into downstream sociotechnical impacts. This may help ensure regulators are less easily dazzled by the "newness" of new AI applications themselves (Mandel, 2017), and enable them to become more aware of how different analogies or metaphors used by the public and lawmakers can highlight different regulatory narratives, in potentially counterproductive ways (Crootof & Ard, 2021).

Another policy could involve (2) setting up a cross-ecosystem agency to focus on "legal foresighting" (Laurie et al., 2012), and forecasting methodologies (Avin, 2019; Karnofsky 2021) aimed at eliciting AI's applications' disparate sociotechnical impacts, link these to potential and actual regulatory rationales, and study the shifting material textures and problem logics around that application. In particular, this can support more democratic and inclusive stakeholder debate over the choices affected parties would seek to make around the deployment of potentially disruptive AI breakthrough capabilities (Cremer & Whittlestone, 2021).

Regulatory tailoring and scope

Relatedly, the sociotechnical change-centric lens helps in *tailoring* regulatory solutions to effective clusters of AI techniques, applications, users, and societal effects. Rather than consign regulators to confront self-similar AI challenges (e.g., around meaningful human control, susceptibility to adversarial attack, unaccountable opacity in algorithmic decision-making) many times across individual legal domains (Crootof & Ard, 2021, p. 1), this approach highlights common themes, underlying material value chains, or shared usage problem logics of AI, as they are expressed across these various domains.

In practical terms, improved regulatory tailoring could involve (3) the establishment of various institutions and 'meta-regulatory' oversight mechanisms—connoting "activities occurring in a wider regulatory space, under the auspices of a variety of institutions, including the state, the private sector and public interest groups [which] may operate in concert or independently" (Grabosky, 2017, p. 150). Such mechanisms could foster improved cross-regime dialogue of AI policy (Cihon et al., 2020b). This can support regulatory harmonization or the bundling of regulatory interventions for various AI applications, where appropriate. It can also examine how and where different regimes and institutions can exploit the same regulatory levers (e.g., compute hardware production) that intersect with the AI development value chain.

Another solution to improve adequate sociotechnical scoping of regulation could be found in (4) newly-established mechanisms and fora for dialogue amongst actors in the AI space, including not just direct developers and policymakers, but the broad set of actors that have a hand in shaping the "problem logic"—in terms of the problem's origins, contributing factors, or barriers to regulation. The aim of such discussions would be to reconfigure some of these wider conditions or parameters to be more supportive or conducive to AI regulation, or—where they are not very tractable—to explore alternative levers or vectors for regulation. Finally, it can promote the exchange of best practices or lessons learned around how regulators can address some of the problem logics that generate the problem or impede regulation.

Regulatory timing

In terms of regulatory *timing and responsiveness*, a study of AI's sociotechnical changes highlights the inadequacies of governance strategies that are grounded in either an attempt to predict sociotechnical changes in detail, or in reactive responses which prefer to "wait out" technological change until its societal impact has become adequately clear. The former is strained—though not rendered impossible—by the fact that accurate prediction of sociotechnical impacts is difficult; the latter strategy has the flaw that it demands a high threshold of epistemic clarity about a technology's impacts before acting, when in fact such clarity is rarely achieved, even decades after a technology's deployment (Horowitz, 2020). By contrast, the sociotechnical approach emphasizes the importance of anticipatory and adaptive regulatory approaches (Maas, 2019b). This could be implemented by (5) incorporating provisions such as sunset clauses that prompt re-examination (at the domestic level) or designating authoritative interpreters (at the international level), amongst many other avenues. Such provisions can help mitigate some of the information problems (about both the present and future) facing AI regulation, by ensuring such regulation can remain adaptive and "scalable" to ongoing change in both AI technology and its sociotechnical effects.

Regulatory design

Finally, in terms of *regulatory design*, the sociotechnical change lens highlights when and why governance ought to prefer technology-neutral rules, and where they should pursue technology-specific rules. By considering the specific regulatory or governance rationale in play, we may understand when or whether technological neutrality is to be preferred. As imperfect general heuristic, Bennett-Moses (2017, p. 586) argues that "regulatory regimes should be technology—neutral to the extent that the regulatory rationale is similarly neutral." In this view, the point is not to find a regulatory strategy that already details long lists of anticipated future applications of AI. The idea is rather to develop institutional mechanisms that are up to the task of managing distinct problem logics—new ethical challenges, security threats, safety risks, structural shifts, opportunities for benefit, or governance disruptions—in a way that can be relatively transferable across, or agnostic to, the specific AI techniques used to achieve those affects. In practice, this may involve (6) clearer guidelines about the circumstances in which AI regulations these should rely on standards or rules, and when they should be tech-specific or tech-neutral, as well as retroactive assessments of when and how different regulatory instruments fare over time.

Conclusion

This chapter has introduced, articulated, and evaluated a "sociotechnical change-centric" perspective on aligning AI regulation.

It first briefly sketched the general value of a "change-centric" approach to the problems facing the AI governance ecosystem. It next articulated a framework focused on Bennett Moses's account of regulation for sociotechnical change. It explored how, when, and why laws and regulations for AI ought to tailor themselves to broad *sociotechnical change,* rather than local *technological change*. It applied this model to AI technology to show how this model allows a better connection of AI applications to five forms of sociotechnical change—and how these in turn can be mapped to six types of regulatory *rationales*—market failures, risks to human health or safety, risks to moral interests, rights or values, new threats to social solidarity or the democratic process, or new threats to the coherence, efficacy or integrity of the existing legal system.

The chapter then turned to the mirror question of how, having established a need for AI governance in a specific case, regulators might craft policy interventions to the particular regulatory target of AI. This involved a consideration of both the material textures of AI applications, but especially demands focus on the "problem logics" involved. It argued that socio-technical changes created by AI applications can be disambiguated into six specific types of challenges—ethical challenges, security threats, safety risks, structural shifts, public goods, and governance disruption—which come with distinct problem features (origins, contributory factors, barriers to regulation), and which each are susceptible to (or demand) different governance responses.

Finally, the chapter concluded by sketching some indicative institutional and regulatory actions that might draw on this framework to improve regulatory *triage*, *tailoring*, *timing and responsiveness*, and *design* of AI policy.

To be clear, an emphasis on sociotechnical change is not an unprecedented insight in scholarship on law, regulation, and new technology. However, in a fragmented and incipient AI governance landscape, it remains a valuable tool, that deserves greater attention and use. As such, "sociotechnical change" should be considered not a new or substitute paradigm for AI governance, but rather a complementary perspective. Such a lens is subject to its own conditions and limits, but when used cautiously, can offer regulators a more considered understanding of which of AI's challenges are possible, plausible, or already-pervasive—and how these might be best met.

Acknowledgements

For valuable comments and feedback on early drafts, I thank Jess Whittlestone, Jack Clark, Harry Surden, and the editors, especially our section editors Valerie Hudson and Justin Bullock. I also thank Christina Korsgaard for her support throughout the writing process. Any remaining errors are all my own.

Note

1. An earlier version of this framework is presented and unpacked in further detail in (Maas 2020, pp. 166–186). Note, this version referred to the "public goods" as the "common goods" problem logic, instead.

References

Amodei, Dario, Olah, Chris, Steinhardt, Jacob, Christiano, Paul, Schulman, John, & Mané, Dan. (2016). Concrete problems in AI safety." *arXiv:1606.06565 [cs]*. http://arxiv.org/abs/1606.06565.

Ard, B. J., & Crootof, Rebecca. (2020). The case for "technology law". *Nebraska Governance & Technology Center*. https://ngtc.unl.edu/blog/case-for-technology-law.

Avin, Shahar. (2019). Exploring artificial intelligence futures. *AIHumanities*. http://aihumanities.org/en/journals/journal-of-aih-list/?board_name=Enjournal&order_by=fn_pid&order_type=desc&list_type=list&vid=15.

Balkin, Jack M. (2015). The path of robotics law. *California Law Review Circuit* 6, 17.

Beaumier, Guillaume, Kalomeni, Kevin, Campbell-Verduyn, Malcolm, Lenglet, Marc, Natile, Serena, Papin, Marielle, Rodima-Taylor, Daivi, Silve, Arthur, & Zhang, Falin. (2020). Global regulations for a digital economy: Between new and old challenges. *Global Policy 11* (4), 515–522. https://doi.org/10.1111/1758-5899.12823.

Belfield, Haydn. (2020, February 12). Activism by the AI community: Analysing recent achievements and future prospects. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, 15–21. http://dl.acm.org/doi/10.1145/3375627.3375814.

Bennett Moses, Lyria. (2007a). Recurring dilemmas: The law's race to keep up with technological change. *University of New South Wales Faculty of Law Research Series 21* , 1–81. http://www.austlii.edu.au/au/journals/UNSWLRS/2007/21.html.

Bennett Moses, Lyria. (2007b). Why have a theory of law and technological change? *Minnesota Journal of Law, Science & Technology 8* (2), 589–606.

Bennett Moses, Lyria. (2013). How to think about law, regulation and technology: Problems with 'technology' as a regulatory target. *Law, Innovation and Technology 5* (1), 1–20.

Bennett Moses, Lyria. (2017). Regulating in the face of sociotechnical change. In Roger Brownsword, Eloise Scotford, & Karen Yeung (Eds.), *The Oxford handbook of law, regulation, and technology* (pp. 573–596). Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49.

Beyleveld, Deryck, & Brownsword, Roger. (2012). Emerging technologies, extreme uncertainty, and the principle of rational precautionary reasoning. *Law, Innovation & Technology 4* (1), 35–65.

Brownsword, Roger. (2019). Law disrupted, law re-imagined, law re-invented. *Technology and Regulation 1* (1), 10–30.

Brownsword, Roger, Scotford, Eloise, & Yeung, Karen. (2017). Law, regulation, and technology: The field, frame, and focal questions. In Roger Brownsword, Eloise Scotford, & Karen Yeung (Eds.), *The Oxford handbook of law, regulation and technology*. (pp. 3–38). Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-1.

Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn, Krueger, Gretchen, Hadfield, Gillian, Khlaaf, Heidy, Yang, Jingying, Toner, Helen, Fong, Ruth, Maharaj, Tegan, Koh, Pang Wei, Hooker, Sara, Leung, Jade, Trask, Andrew, Bluemke, Emma, Lebensold, Jonathan, O'Keefe, Cullen, Koren, Mark, . . . Anderljung, Markus. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv:2004.07213 [cs]*. http://arxiv.org/abs/2004.07213.

Buiten, Miriam C. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation 10* (1), 41–59.

Burri, Thomas. (2017). International law and artificial intelligence. *German Yearbook of International Law 60* , 91–108.

Calo, Ryan. (2015). Robotics and the lessons of cyberlaw. *California Law Review 103* , 513–564.

Cihon, Peter. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. Technical Report. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Cihon, Peter, Kleinaltenkamp, Moritz J., Schuett, Jonas, & Baum, Seth D. (2021). AI CERTIFICATION: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society 2* (4), 200–209.

Cihon, Peter, Maas, Matthijs M., & Kemp, Luke. (2020a). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy 11* (5), 545–556.

Cihon, Peter, Maas, Matthijs M., & Kemp, Luke. (2020b.) Should artificial intelligence governance be centralised? Design lessons from history. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, 228–234. http://dl.acm.org/doi/10.1145/3375627.3375857.

Clark, Jack. (2024). Information markets and AI development. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Clark, Jack, & Hadfield, Gillian K. (2019). Regulatory markets for AI safety. *arXiv:2001.00078 [cs.CY]*. https://arxiv.org/abs/2001.00078

Clark, Jack, & Perrault, Ray. (2023). Artificial Intelligence Index Report 2023. Stanford Institute for Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

Cremer, Carla Zoe, & Whittlestone, Jess. (2021). Artificial canaries: Early warning signs for anticipatory and democratic governance of AI. *International Journal of Interactive Multimedia and Artificial Intelligence 6* (5), 100–109.

Crootof, Rebecca. (2019). Jurisprudential space junk: Treaties and new technologies. In Chiara Giorgetti & Natalie Klein (Eds.), *Resolving conflicts in the law* (106–129). Brill. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.

Crootof, Rebecca, & Ard, B. J. (2021). Structuring techlaw. *Harvard Journal of Law & Technology 34* (2), 347–417. https://jolt.law.harvard.edu/assets/articlePDFs/v34/1.-Crootof-Ard-Structuring-Techlaw.pdf.

Cummings, Mary L., Roff, Heather, Cukier, Kenneth, Parakilas, Jacob, & Bryce, Hannah. (2018). *Artificial intelligence and international affairs: Disruption anticipated*. Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Cussins, Jessica. (2020). National and international AI strategies. *Future of Life Institute*. https://futureoflife.org/national-international-ai-strategies/.

Dafoe, Allan. (2020). AI governance: Opportunity and theory of impact. https://www.allandafoe.com/opportunity.

Dunlap, Charles. (2008). Lawfare Today: A Perspective. *Yale Journal of International Affairs Winter* 146–154.

Erdelyi, Olivia J., & Goldsmith, Judy. (2018). Regulating artificial intelligence: Proposal for a global solution. *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society* (pp. 95–101). http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf.

Feijóo, Claudio, Kwon, Youngsun, Bauer, Johannes M., Bohlin, Erik, Howell, Bronwyn, Jain, Rekha, Potgieter, Petrus, Vu, Khuong, Whalley, Jason, & Xia, Jun. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy 44* (6), 101988. https://doi.org/10.1016/j.telpol.2020.101988.

Feldstein, Steven (2023). The consequences of generative AI for democracy, governance and war. *Survival* 65 (5), 117–142. https://doi.org/10.1080/00396338.2023.2261260.

Fischer, Sophie-Charlotte, Leung, Jade, Anderljung, Markus, O'Keefe, Cullen, Torges, Stefan, Khan, Saif M., Garfinkel, Ben, & Dafoe, Allan. (2021). AI policy levers: A review of the U.S. government's tools to shape AI research, development, and deployment. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf.

Fjeld, Jessica, Hilligoss, Hannah, Achten, Nele, Daniel, Maia Levy, Feldman, Joshua, & Kagay, Sally. (2019). *Principled artificial intelligence: A map of ethical and rights-based approaches*. Berkman Klein Center for Internet & Society at Harvard University. https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf.

Floridi, Luciano, Cowls, Josh, Beltrametti, Monica, Chatila, Raja, Chazerand, Patrice, Dignum, Virginia, Luetge, Christoph, Madelin, Robert, Pagallo, Ugo, Rossi, Francesca, Schafer, Burkhard, Valcke, Peggy, & Vayena, Effy. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines 28* (4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Friedman, David D. (2001). Does technology require new law? *Public Policy 71* , 16.

Gabriel, Iason. (2020). Artificial intelligence, values, and alignment. *Minds and Machines 30* (3), 411–437.

Garfinkel, Ben, & Dafoe, Allan. (2019). How does the offense–defense balance scale? *Journal of Strategic Studies 42* (6), 736–763.

Grabosky, Peter. (2017). Meta-Regulation. In Peter Drahos (Ed.), *Regulatory theory, foundations and applications* (pp. 149–162). ANU Press. https://www.jstor.org/stable/j.ctt1q1crtm.17.

Gruetzemacher, Ross, & Whittlestone, Jess. (2022). The transformative potential of artificial intelligence. *Futures 135* , *102884*. https://www.sciencedirect.com/science/article/pii/S0016328721001932

Guihot, Michael, Matthew, Anne F., & Suzor, Nicolas. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt Journal of Entertainment & Technology Law*. https://papers.ssrn.com/abstract=3017004.

Gutierrez, Carlos Ignacio, & Marchant, Gary. (2021). *A global perspective of soft law programs for the governance of artificial intelligence*. Sandra Day O'Connor College of Law, Arizona State University. https://lsi.asulaw.org/softlaw/the-report/.

Gutierrez, Carlos Ignacio, Marchant, Gary E., & Tournas, Lucille. (2020). Lessons for artificial intelligence from historical uses of soft law governance. *JURIMETRICS 61* (1), 1–18. https://papers.ssrn.com/abstract=3775271.

Hernandez-Orallo, Jose, Martınez-Plumed, Fernando, Avin, Shahar, Whittlestone, Jess, & Ó hÉigeartaigh, Seán. (2020). AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues. *European Conference on Artificial Intelligence 8* . https://ecai2020.eu/papers/1364_paper.pdf.

Hopster, Jeroen K. G., & Maas, Matthijs M. (2023). The technology triad: Disruptive AI, regulatory gaps and value change. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00305-5.

Horowitz, Michael C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power.

Horowitz, Michael C. (2020). Do emerging military technologies matter for international politics? *Annual Review of Political Science 23* (1), 385–400.

ITU. (2019). AI for Good Global Summit 2019 insights. https://itu.foleon.com/itu/aiforgood2019/home/.

Jelinek, Thorsten, Wallach, Wendell, & Kerimi, Danil. (2020). Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. AI and Ethics *1*, 141–150. https://doi.org/10.1007/s43681-020-00019-y.

Jobin, Anna, Ienca, Marcello, & Vayena, Effy. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1* , 389–399.

Karnofsky, Holden (2021). AI timelines: where the arguments, and the "experts," stand. *Cold Takes*. https://www.cold-takes.com/where-ai-forecasting-stands-today/

Kemp, Luke, Cihon, Peter, Maas, Matthijs Michiel, Belfield, Haydn, Cremer, Zoe, Leung, Jade, & Ó hÉigeartaigh, Seán. (2019). UN high-level panel on digital cooperation: A proposal for international AI governance. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

Krakovna, Victoria, Uesato, Jonathan, Mikulik, Vladimir, Rahtz, Matthew, Everitt, Tom, Kumar, Ramana, Kenton, Zachary, Leike, Jan, & Legg, Shane. (2020). Specification gaming: The flip side of AI ingenuity. *Deepmind* (blog). https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

Kumar, Ram Shankar Siva, O. Brien, David, Albert, Kendra, Viljöen, Salomé, & Snover, Jeffrey. (2019). Failure modes in machine learning systems. *arXiv:1911.11034 [cs, stat]*. http://arxiv.org/abs/1911.11034.

Kunz, Martina, & ÓhÉigeartaigh, Seán. (2021). Artificial intelligence and robotization. In Robin Geiss & Nils Melzer (Eds.), *The Oxford handbook of the international law of global security*. (pp. 624–640). Oxford University Press. https://papers.ssrn.com/abstract=3310421.

Laurie, Graeme, Harmon, Shawn H. E., & Arzuaga, Fabiana. (2012). Foresighting futures: Law, new technologies, and the challenges of regulating for uncertainty. *Law, Innovation & Technology 4* (1), 1–33.

Law Library of Congress. (2019). *Regulation of artificial intelligence in selected jurisdictions*. The Law Library of Congress. https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf.

Leike, Jan, Martic, Miljan, Krakovna, Victoria, Ortega, Pedro A., Everitt, Tom, Lefrancq, Andrew, Orseau, Laurent, & Legg, Shane. (2017). AI safety gridworlds. *arXiv:1711.09883 [cs]*. http://arxiv.org/abs/1711.09883.

Leung, Jade. (2019). Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies. University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665.

Liu, Hin-Yan, Maas, Matthijs, Danaher, John, Scarcella, Luisa, Lexer, Michaela, & Van Rompaey, Leonard. (2020). Artificial intelligence and legal disruption: A new model for analysis. *Law, Innovation and Technology 12* (2), 205–258. https://doi.org/10.1080/17579961.2020.1815402.

Liu, Hin-Yan, & Maas, Matthijs M. (2021). "Solving for X?" Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence. *Futures 126* , 22.

Lorenz, Philippe. (2020). *AI governance through political fora and standards developing organizations*. Stiftung Neue Verantwortung. https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations.

Maas, Matthijs M. (2018). Regulating for "normal AI accidents": Operational lessons for the responsible governance of artificial intelligence deployment." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New York, Association for Computing Machinery, pp. 223–228. https://doi.org/10.1145/3278721.3278766.

Maas, Matthijs M. (2019a). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy 40* (3), 285–311.

Maas, Matthijs M. (2019b.) Innovation-proof governance for military AI? How I learned to stop worrying and love the bot. *Journal of International Humanitarian Legal Studies 10* (1), 129–157.

Maas, Matthijs M. (2019c). International law does not compute: Artificial intelligence and the development, displacement or destruction of the global legal order. *Melbourne Journal of International Law 20* (1), 29–56.

Maas, Matthijs M. (2020). Artificial intelligence governance under change: Foundations, facets, frameworks. University of Copenhagen. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3833395.

Mandel, Gregory N. (2017). Legal evolution in response to technological change. In Roger Brownsword, Eloise Scotford, & Karen Yeung, (Eds.), *The Oxford handbook of Law, regulation and technology*. (pp. 225–245). Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-45.

Marchant, Gary E. (2011). The growing gap between emerging technologies and the law. In Gary E. Marchant, Braden R. Allenby, & Joseph R. Herkert (Eds.), *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem* (pp. 19–33). The International Library of Ethics, Law and Technology. Springer Netherlands. https://doi.org/10.1007/978-94-007-1356-7_2.

Newman, Jessica Cussins. (2020). *Decision points in AI governance*. Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf.

Norman, Donald A. (2013). *The design of everyday things*, revised and expanded edition. Basic Books.

ÓhÉigeartaigh, Seán S., Whittlestone, Jess, Liu, Yang, Zeng, Yi, & Liu, Zhe. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology 33* , 571–593. https://doi.org/10.1007/s13347-020-00402-x.

O'Keefe, Cullen, Cihon, Peter, Garfinkel, Ben, Flynn, Carrick, Leung, Jade, & Dafoe, Allan. (2020). The windfall clause: Distributing the benefits of AI for the common good. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, pp. 327–331. https://doi.org/10.1145/3375627.3375842.

Pasquale, Frank. (2019). The second wave of algorithmic accountability. *LPE Project*. https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/.

Petit, Nicolas. (2017). *Law and regulation of artificial intelligence and robots—conceptual framework and normative implications*. Social Science Research Network (SSRN) Scholarly Paper. https://papers.ssrn.com/abstract=2931339.

Prosser, Tony. (2010). The *r*egulatory *e*nterprise: Government, *r*egulation, and *l*egitimacy. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199579839.001.0001/acprof-9780199579839.

Prunkl, Carina E. A., Ashurst, Carolyn, Anderljung, Markus, Webb, Helena, Leike, Jan, & Dafoe, Allan. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence 3* (2), 104–110. https://doi.org/10.1038/s42256-021-00298-y.

Rahwan, Iyad, Cebrian, Manuel, Obradovich, Nick, Bongard, Josh, Bonnefon, Jean-François, Breazeal, Cynthia, Crandall, Jacob W., Christakis, Nicholas A., Couzin, Iain D., Jackson, Matthew O., Jennings, Nicholas R., Kamar, Ece, Kloumann, Isabel M., Larochelle, Hugo, Lazer, David, McElreath, Richard, Mislove, Alan, Parkes, David C., Pentland, Alex 'Sandy', . . . Wellman, Michael. (2019). Machine behaviour. *Nature 568* (7753): 477. https://doi.org/10.1038/s41586-019-1138-y.

Raji, Inioluwa Deborah, Smart, Andrew, White, Rebecca N., Mitchell, Margaret, Gebru, Timnit, Hutchinson, Ben, Smith-Loud, Jamila, Theron, Daniel, & Barnes, Parker. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, Barcelona, Association for Computing Machinery, pp. 33–44. https://doi.org/10.1145/3351095.3372873.

Roff, Heather M. (2018). Advancing human security through artificial intelligence. In M. L. Cummings, Heather M. Roff, Kenneth Cukier, Jacob Parakilas, & Hannah Bryce (Eds.), *Artificial intelligence and international affairs: Disruption anticipated* (pp. 19–28). Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Russell, Stuart, & Norvig, Peter. (2016). *Artificial intelligence: A modern approach*. 3rd ed. Pearson.

Sapignoli, Maria. (2021). The mismeasure of the human: Big data and the "AI turn" in global governance. *Anthropology Today 37* (1), 4–8.

Scherer, Matthew U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology 29* (2), 354–398. http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf.

Schiff, Daniel, Biddle, Justin, Borenstein, Jason, & Laas, Kelly. (2020). What's next for AI ethics, policy, and governance? A global overview. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, pp. 153–58. http://dl.acm.org/doi/10.1145/3375627.3375804.

Schuett, Jonas. (2019). A legal definition of AI. *arXiv:1909.01095 [cs]*. http://arxiv.org/abs/1909.01095.

Severance, C. (2016). Bruce Schneier: The security mindset. *Computer 49* (2), 7–8.

Shevlane, Toby, & Dafoe, Allan. (2020). The offense–defense balance of scientific knowledge: Does publishing AI research reduce misuse?" *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, New York, Association for Computing Machinery, pp. 173–79. https://doi.org/10.1145/3375627.3375815.

Smith, Bryant Walker. (2020). New technologies and old treaties. *AJIL Unbound 114* , 152–157.

Stahl, Bernd Carsten, Antoniou, Josephina, Ryan, Mark, Macnish, Kevin, & Jiya, Tilimbe. (2021). Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01148-6.

Stix, Charlotte. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics 27* (1), 15.

Stone, Peter, Brooks, Rodney, Brynjolfsson, Erik, Calo, Ryan, Etzioni, Oren, Hager, Greg, Hirschberg, Julia, Kalyanakrishnan, Shivaram, Kamar, Ece, Kraus, Sarit, Leyton-Brown, Kevin, Parkes, David, Press, William, Saxenian, AnnaLee, Shah, Julie, Tambe, Milind, & Teller, Astro. (2016). *Artificial intelligence and life in 2030*. Stanford University. http://ai100.stanford.edu/2016-report.

Trajtenberg, Manuel. (2018). *AI as the next GPT: A political-economy perspective*. National Bureau of Economic Research. Working Paper. http://www.nber.org/papers/w24245.

Turner, Jacob. (2018). *Robot rules: Regulating artificial intelligence*. Springer Berlin Heidelberg.

van der Loeff, Agnes Schim, Bassi, Iggy, Kapila, Sachin, & Gamper, Jevgenij. (2019). AI ethics for systemic issues: A structural approach. *arXiv:1911.03216 [cs.AI]*. http://arxiv.org/abs/1911.03216 (January 13, 2020).

Verbruggen, Maaike. (2020). In defense of technological determinism." *7th European Workshop in International Studies*. Brussels, Belgium.

Watts, Sean. (2015). Regulation-tolerant weapons, regulation-resistant weapons and the law of war. *International Law Studies 91* , 83.

Weidinger, Laura, Rauh, Maribeth, Marchal, Nahema, Manzini, Arianna, Hendricks, Lisa Anne, Mateos-Garcia, Juan, Bergman, Stevie, Kay, Jackie, Griffin, Conor, Bariach, Ben, Gabriel, Iason, Rieser, Verena, Isaac, William. (2023). Sociotechnical Safety Evaluation of Generative AI Systems. arXiv.https://doi.org/10.48550/arXiv.2310.11986.

Whittlestone, Jess, & Clark, Jack (2021). Why and how governments should monitor AI development. *arXiv:2108.12427 [cs]*. http://arxiv.org/abs/2108.12427

Winter, Christoph, Schuett, Jonas, Martínez, Eric, Van Arsdale, Suzanne, Araújo, Renan, Hollman, Nick, Sebo, Jeff, Stawasz, Andrew, O'Keefe, Cullen, & Rotola, Giuliana. (2021). *Legal Priorities Research: A Research Agenda*. Legal Priorities Project. https://www.legalpriorities.org/research_agenda.pdf.

Zhang, Baobao, & Dafoe, Allan. (2020). U.S. public opinion on the governance of artificial intelligence." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, pp. 187–93. http://dl.acm.org/doi/10.1145/3375627.3375827.

Zhang, Daniel, Mishra, Saurabh, Brynjolfsson, Erik, Etchemendy, John, Ganguli, Deep, Grosz, Barbara, Lyons, Terah, Manyika, James, Niebles, Juan Carlos, Sellitto, Michael, Shoham, Yoav, Clark, Jack, & Perrault, Raymond. (2021). *Artificial intelligence index report 2020*. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf.

Zwetsloot, Remco, & Dafoe, Allan. (2019). Thinking about risks from AI: Accidents, misuse and structure. *Lawfare*. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

SECTION IV

FRAMEWORKS AND APPROACHES FOR AI GOVERNANCE

YU-CHE CHEN AND MATTHEW M. YOUNG

CHAPTER 19

THE CHALLENGE OF AI GOVERNANCE FOR PUBLIC ORGANIZATIONS

JUSTIN B. BULLOCK, HSINI HUANG, KYOUNG-CHEOL KIM, AND MATTHEW M. YOUNG

Introduction

Government and public organizations have already begun to rely extensively on artificial intelligence (AI) systems to provide public services (Young et al., 2019), although adoption has been uneven (Bullock et al., 2020). Meanwhile, even though many public service providers have co-evolved along with these assistive technologies (Bovens & Zouridis, 2002), their evolution has varied across both agency and type of task (Busch & Henriksen, 2018). In the midst of this real-time process, which is birthing new forms of public organizations, we argue that it is useful to revisit their classic, complementary formulations to examine how this co-evolution with technology challenges and otherwise affects questions of modern governance generally, and AI governance in particular.

To this end, we return to Max Weber's concept of *ideal type bureaucracy* (Weber, 1922) and Herbert Simon's description of *administrative behavior* (Simon, 1997). Given the increasing capability of AI to act as agents within public organizations (Bullock, 2019; Bullock & Kim, 2020), we maintain that Weber's characterization introduces three challenges, including: (1) a decreased scope of tasks for human bureaucrats, (2) a loss in human managerial control, and (3) an evolving bureaucracy structure. The framework provided by Simon's version highlights several problems as well, including: (1) a reshaping of the bounded rationality of agents, (2) value alignment of public organizations, and (3) a changing balance of formal and informal communication.

Finally, we argue that considering AI challenges through these classic views reveals two very troubling trends for public organizations and their behavior: increases in dehumanization and losses in human control. At the same time, reflecting upon Weber and Simon can help guide the field through these two dangers, which were core concerns to both writers.

Bounded rationality of agents, hierarchy, delineation of facts and values, specialization, and human judgment are offered as tools by which modern public organizations can reassert human dignity and human control.

This chapter will first introduce the nature of AI-related risks in the public sector. From there, we review Weber and Simon's theoretical approaches to understanding administrative functioning, structure, and behaviors in relation to the use of AI. Finally, in response to the pervasive spread of AI across the decision-making functions of many organizations and governing systems, we propose strategies to address these threats to human influence and control.

Artificial Intelligence and Public Organizations

Machine learning-based AI technology has rapidly advanced over the past two decades. As its ability to successfully optimize decision functions under complex scenarios increases, some forecast that AI credibly threatens to displace "white collar" labor, including professional roles found in public organizations. In the U.S. context, early signalers of intended AI adoption include the U.S. General Services Administration (GSA), which plans to use AI for rule setting at the institutional level (Farber et al., 2021). The U.S. Department of Defense is developing a similar rule-setting application, the Joint All-Domain Command and Control (JADC2), and is also designing and conducting trials for autonomous or "self-flying/navigating" airplanes and ships (Hambling, 2021; Dempsey, 2021). In most cases, the stated motivation behind the push for adopting AI is to increase the organization's efficiency.

These systems, however, are not planned, adopted, or implemented in a vacuum, but rather exist within larger organizational contexts and frameworks. Public organizations operate under particular pressures to optimize along multiple value dimensions, some of which, such as transparency, accountability, and equity, often require tradeoffs against efficiency. As a technical artifact, AI's behavior is conditioned by its inner environment and interfaces (Simon, 1997), leading to poorly understood risks to the entire decision-making structure. In most organizations, and virtually all public ones, this structure is commonly understood to be bureaucratic.

Bureaucracies structure authority to act across clear vertical and horizontal lines of command, and employ specialists whose jobs consist of purpose-clustered tasks and responsibilities (Weber, 1922; Waters & Waters, 2015; Simon, 1946, 1997). Rational legal authority is thus derived from hierarchical relations between agents, with individual agents responsible for explicitly defined, specialized sets of tasks. This structure has its limits and tradeoffs. On one hand, the professionalization of bureaucrats—those dedicated specialists protected from political threats that Weber called *Beamte*—demands standardization, rationality, control, and administrative behavior that reflects the hierarchical order and will of the authority. On the other hand, formal rules, laws, and delegation structures are designed in a way that requires organizational agents to practice some level of discretion (Rourke, 1972).

Indeed, the scope and complexity of tasks necessary to solve collective-action problems requires delegation of work to agents, who in turn must exercise discretion in deciding how to perform them (Arrow, 1984; Bertelli, 2012; Huber & Shipan, 2002). Prior work on public sector innovation via information and communication technology (ICT) has produced three related theories tracking ICT's effect on administrative discretion: systems-level bureaucracy (Bovens & Zouridis, 2002), digital discretion (Busch & Henriksen, 2018; Busch & Eikebrokk, 2019), and artificial discretion (Young et al., 2019). In this framework, artificial discretion extends the general concept of digital discretion to describe machine learning-based AI.

AI agents are designed to practice discretion in decision making, either alone (i.e., the task is fully automated using AI) or as a supporting player (i.e., as a second set of eyes, akin to a human peer or collaborator). Take, for example, the use of AI for facial recognition. The desired outcome of this discrete task is a determination of whether a given facial image can be reliably paired to a separate image with associated metadata about the individual (e.g., name, age, known residence, etc.). The AI-as-agent ultimately produces a decision: it reports a credible match (or a set of possible matches) or reports that it has failed to do so. On a sub-task level, this process involves comparing n pre-known images of faces against the new image; at an even lower level, the AI agent is performing k subroutines, evaluating and comparing the geometries, shade gradients, and other constituent elements of the images.

This process, however, is not "rule bound." The AI agent's final decision cannot be deterministically reverse engineered from the set of input factors: many of the most "sophisticated" and popular machine-learning (ML) architectures (e.g., artificial neural networks) involve stochastic processes, just as with a human agent. While it is possible to articulate and enumerate certain steps, subject to the limitations of time and physics, any given agent's final decision output is probabilistic, not deterministic. As such, from an organizational perspective, AI agents are afforded the same discretion for this task as any human agent: they can be trained, and held to performance standards, but the outcome of any discrete task is not perfectly knowable a priori. In this way, these AI systems are acting as artificial bureaucrats empowered with artificial discretion (Young et al., 2021; Bullock & Kim, 2020).

As with human bureaucrats, this power carries an implicit risk of error. AI has been shown to fail to generate precise outcomes, especially in novel circumstances (Bellotti, 2021). Other work has found that AI cannot perfectly secure and process data and information while simultaneously developing and adjusting its protocol, resulting in poor decision making (Bullock & Kim, 2020). Autonomous vehicles, reliable in repetitive, predictable situations, may judge and react sub-optimally when faced with unexpected road, weather, or pedestrian conditions. Minimizing this risk can occur at the individual agent level (upgrading one ML-based system to a higher performer), or at the institutional level (structurally changing whether and how to integrate the AI into the organization's decision-making and task-executing processes).

These two distinct opportunities for managing AI error (individual vs. institution) introduce a tradeoff to scholars selecting their unit of analysis for studying the benefits and risks of AI in public organizations. Our goal for this chapter is to offer the reader a framework for leveraging two classic perspectives on administration that emphasize each analytic unit in their own way. In the following section, we synthesize the work of Weber and Simon to provide a construct for considering both the systemic and discrete risks and opportunities of AI for public organizations.

Theoretical Lenses: Weber & Simon

Max Weber and Herbert Simon's influential descriptions of bureaucratic functioning and administrative behavior can lay the groundwork for understanding AI's challenges to government. Importantly, we do not see these as competing descriptions of organizational behavior, but as distinct characterizations driven by differing goals and emphases, producing different frames, language, and concepts by which to evaluate their use of AI.

Weber sought to describe the design of a formal, rational structure to produce a controlled, standardized public organization, reliant on professional agents to adapt the delivery of public services to the needs of any particular situation. He also speculated on the troubling parallels between his description of bureaucratic structure and that of machine efficiency. Water & Waters (2015) identified six guiding features of the ideal type of bureaucracy:

1. Well-defined and organized competencies.
2. A hierarchical organizational structure, with codified channels for making and transmitting information and decisions.
3. Modern administration is based on (a) documents preserved as original copies or concepts, and (b) a staff of subordinated *Beamte* and writers of all kinds. *Beamte* working in the office, together with relevant resources including material goods and documents, constitute a bureau.
4. The work of the *Beamte* usually requires significant training for the specific tasks to be accomplished.
5. A full-fledged position occupies all the professional energy of the *Beamte* to process its tasks, regardless of limits to his mandatory working hours.
6. The duties of the position undertaken by the *Beamte* are based on general learnable rules and regulations, which are more or less firm and more or less comprehensible. The knowledge of these rules and regulations thus constitutes a special kind of 'applied science,' which the *Beamte* possesses.

Simon, on the other hand, started from the perspective of the individual decision, individual decision maker, and the general challenges of organizing information, communication, and general problem solving for effective and efficient organizational behavior. Simon focused on information flows across an organization, and how to arrange these flows so that the decision maker with the best information, most representative of the values of the organization, addresses a given problem (Simon, 1997). Simon also believed that this design could benefit types of artificial systems beyond organizations. His conditions for a curriculum of social design, that is, an identification of the core features that influence social behavior (in large part through organizations), rely on six guiding ideas:

1. *Bounded rationality.* The meaning of rationality in situations where the complexity of the environment is immensely greater than the computational powers of the adaptive system.
2. *Data for planning*. Methods of forecasting the use of prediction and feedback in control.

3. *Identifying the client*. Professional–client relations, society as the client, the client as player in a game.
4. *Organizations in social design*. Not only is social design carried out mainly by people working in organizations, but an important goal of the design is to fashion and change social organizations in particular.
5. *Time and space horizons*. The discounting of time, defining progress, and managing attention.
6. *Designing without final goals.* Designing for future flexibility, design activity as goal, and designing an evolving system. (Simon, 1996, p. 166)

It should further be noted that both Weber and Simon were concerned with the roles that both machines and information communication technology use within human organizations, particularly those charged with providing the public good. For example, Weber (1978, p. 979) argues:

> The idea that there can be a law without loopholes is generally strongly contested. Also, the idea that the modern magistrate is nothing more than a "judging machine" is rejected with disgust. In such a judging machine, files and costs would be thrown into the top of the machine in order that it would spit out the verdict along with a mechanical reasoning at the bottom . . . In fact, though, as within the domain of the findings of justice, there are areas where the bureaucratic judge is instructed by the legislature to use "individual" paths to find justice. Moreover, in the area of the actual administrative functions, including all government activities that do not belong to the area of creating or implementing the law, or the finding of justice, the freedom and domination (Herrschaft) of individual approaches to tasks are simply taken for granted. In contrast to such emphases on individualized approaches stand the invariable norms that play a negative role because they restrict the never-to-be-regulated "creative" work of the Beamte, which is normally seen as a positive thing.

In the fourth edition of his classic work *Administrative Behavior* (Simon 1997, p. 22), Simon addresses the role of information communication technologies for administrative behavior, declaring that:

> We have learned by now that the computer, too, is something far different from an oversized adding machine, and far more significant for our society. But its significance is only just beginning to emerge, the appearance of the personal computer about a decade ago perhaps being a decisive turning point. One way to conjecture what important novel tasks computers may take on is to review the many metaphors that have been applied to them. **First, the computer is an incredibly powerful number cruncher**. We have already proceeded, especially in engineering and science, a considerable way toward discovering what can be done by number crunching, but we will find new uses as computer power continues to increase. **Second, the computer is a large memory**, and we are just beginning to explore (for example, on the World Wide Web) how large data bases must be organized so that they can be accessed selectively and cheaply in order to extract information they contain that is relevant to our specific tasks. **Third, the computer is an expert, capable of matching human professional-level performance** in some areas of medical diagnosis, of engineering design, of chess playing, of legal search, and increasing numbers of others. **Fourth, the computer is the core of a new worldwide network of communications**, an "information superhighway." Everyone can now communicate with "everyone," almost instantaneously. **Fifth, the computer is a "giant brain," capable of thinking, problem solving, and yes, making decisions**. We are continually finding new areas of decision—evaluating credit risks, investing funds,

> scheduling factories, diagnosing corporate financial problems—where computers can play an important role or sometimes do the whole task. . . . **The central lesson the computer should teach is that information is no longer scarce** or in dire need of enhanced distribution. In contrast with past ages, we now live in an information rich world.

This focus on information management informs Simon's recommendations on how to match information systems to appropriate technology:

> The key to successful design of information systems lies in matching the technology to the limits of the attentional resources. From this general principle, we can derive several rules of thumb to guide us when we are considering adding a component to an existing information system. In general, an additional component (man or machine) for an information-processing system will improve the system's performance only if:
>
> 1. Its output is small in comparison with its input, so that it conserves attention instead of making additional demands on attention;
> 2. It incorporates effective indexes of both passive and active kinds (indexes are processes that automatically select and filter information for subsequent transmission); and
> 3. It incorporates analytic and synthetic models that are capable not merely of storing and retrieving information, but of solving problems, evaluating solutions, and making decisions. (Simon, 1997, p. 248)

The bureaucracy of Max Weber

The implementation of AI into Weberian bureaucracy presents challenges for the status quo of these organizations, including (1) a decreased scope of tasks for human bureaucrats, (2) the loss of human managerial control, and (3) an evolving bureaucracy.

Decreased scope of tasks for human bureaucrats

The number and extent of AI applications being deployed across public services is slowly, but steadily, increasing. Objective decision making requires rules and rationality, but how might these guidelines, couched in these classic principles of bureaucracy, be reshaped via human–AI collaboration systems?

One major change will be the attenuation of specialization, a crucial element of efficiency, as AI takes over low-end, labor-intensive, low discretionary, and protocol-defined tasks. For example, Agarwal (2018) found machine learning and computer vision technologies to be promising methods for understanding and interpreting images. While scholars disagree on whether disruptive AI innovations will enhance, worsen, or take over the distributed capital to labor (Korinek & Stiglitz, 2018), we predict that human bureaucrats' tasks will gradually focus more on cross-checks and coordination of cross-boundary problems. This result will challenge the specified division of labor scheme proposed by Weber. This uncertainty will soon affect the interdependence of the task process and eventually create inefficiency in public organizations.

Moreover, Weber's ideal type bureaucracy relies on the concept of technocracy, which emphasizes rationality and logic, leading to an interesting juxtaposition of human versus AI technocracy (Sætra, 2020). Before AI's emergence, the terms "technocracy" and "technocrats" indicated rule by experts and their proponents (Putnam, 1977). However, the

entrance of AI complicates the concept within the Weberian organization. Are humans and AI complementary or supplementary in bureaucracy? How might this change the nature of control of bureaucracy? How might it change the structure of bureaucracy as well?

Loss of human managerial control

Weber's view of managerial control of the bureaucracy defines the modern bureaucratic organization as a structure of hierarchy, specialization, and formalization. Hierarchy creates positions and status, and it defines the relationships between supervisors and subordinates. Specialization occurs within this hierarchical structure, as tasks are divided into simple, functional, and routine sub-components. Under this power structure, supervisors can manage both performance and their employees with ease and efficiency. Meanwhile, formalization, critical in large public agencies, supports organizational control by standardizing and documenting the processes by which a task is to be completed. Similarly, for the purpose of control, whether autocratic or monopolistic, Weber advocated minimizing free and arbitrary "bureaucratic judgment," declaring that "for all state activities that fall outside the field of law creation and court procedure, one has become accustomed to claims for the freedom and the paramountcy of individual circumstances" (Weber, 1978, p. 979).

The development of the internet and innovative ICTs improved information flows and communications across both organizations and the globe, but it also radically changed the public management regime (Dunleavy, 2005). Organizations face a growing need for coordination and monitoring to deal with complex, cross-boundary public affairs. Digital governance has spawned new public issues and problems that agencies look to solve using new AI tools and machine algorithms. For example, a national cybersecurity information-sharing platform must develop automated and collaborative data-sharing and reporting protocols to ensure information immediacy and comparability across public and private sectors (Pala & Zhuang, 2019). Japan's Regional Economy and Society Analyzing System (RESAS) uses big data from multiple sources and AI analytics to visualize the performance of the local economy. In this data-driven era of public administration, such processes may generate new, specialized jobs or tasks to coordinate such functions.

This introduction of AI and algorithms challenges the structures, functions, and controls within Weber's ideal bureaucracy by affecting both political legitimacy and managerial capacity (Young et al., 2019; Bullock and Kim, 2020). Organizations must grow and innovate to deal with large and complex public problems, complicating the management of bureaucracy from the perspectives of network (O'Toole, 1997) and complexity theory (Miller & Page, 2007). To prevent the loss of human managerial control, we must cultivate a better understanding of the role of adaptive and self-steering AI agents within the bureaucratic labor force.

Evolving bureaucracy

The implementation of AI is evolving bureaucracy, particularly administrative discretion, presenting another challenge to Weber's bureaucracy. As powerful algorithms affect organizational routines and structures, specialized experts within the ideal bureaucratic organizations may act as information stewards to address growing criticism from the

outside (Huang et al., 2021). Bovens and Zouridis (2002) argue that under the diffusion of ICTs, the roles and functions of public servants have shifted from street- to screen- and system-level bureaucrats, suggesting a more technical administrative shift. They note that new technologies remove or limit the discretion of public servants' professional discretion across multiple tasks, such as communicating with citizens, processing data and information, and making a final judgment.

AI's powerful predictive analysis can deliver a probabilistic-oriented, predictive discretion approach within an organization, directly affecting the knowledge-focus and rule-based practice of administrative personnel. Bullock et al. (2020) present case analyses of policing and public health insurance administration and find the relationship between AI and discretionary power depends on the organizational contexts and task characteristics: system-level changes may not follow in a linear fashion. In policing, for example, tasks often require higher discretion by humans, and citizens and organizations may resist system-level AI transitions that limit that discretion.

The administrative behavior of Herbert Simon

Simon, like Weber, was interested in the question of organizational control, but focused on mechanisms beyond government-controlled organizations that provide public services. Simon (1997) proposed the theory of bounded rationality, based on (1) the human tendency to choose among preferred alternatives, (2) human cognitive limitations, and (3) time-constrained decision making. He considered organizations to be artificial things that systematically process information to comprehend and influence an increasingly complex modern environment.

Human cognitive processing, which Simon explained as having three levels, can be simulated and modeled by tracking information-processing and decision-making tasks. His bounded rationality theory challenged classic rationality thinking and emphasized the limited ability of human beings to adapt, delineate, and act optimally in a complex environment. Simon's insight informs the conclusions of this chapter, that organizational complexity and the radical improvement in artificial algorithms heighten problems of incomplete information about alternatives. Simon's higher mental level requires the most scholarly attention: it involves organizational goals such as problem-solving, concept-attainment, rule induction, and communication processing tasks (Simon, 1979).

Below, we identify three challenges AI presents to organizations based on Simon's conceptualizations of administrative behavior and managing information complexity. We argue that the introduction of AI (1) reshapes the contours of the bounded rationality of agents, (2) complicates the value alignment of public organizations and their mission, and (3) changes the balance of formal and informal communication.

Reshaping the contours of the bounded rationality of agents

Under Simon's information-processing approach to understanding complexity, organizations are the vehicles through which humans can coordinate towards common goals that require multi-agent efforts. Decision-making agents are placed at decision and action

points along the organizational information-processing pathway. Simon (1997) recognized the important role of ICT tools in reshaping the capabilities of individual decision-making agents and the limits placed upon their rationality. Imagine organizations before the advent of personal computers, emails, search engines, and real-time web maps. These new digital communication and processing technologies greatly enhance the decision-making capacities both of modern institutions and the boundedly-rational humans within them.

AI tools have begun to resemble learning and acting agents deployed into organizational decision making, representing a qualitative shift in the nature of the technological tools themselves and their various roles and responsibilities (decision making, communication, and action). This shift cannot help but impact how the organization gathers, analyzes, organizes, and communicates information: human counterparts will require different approaches to management, control, information flow, decision making, and action execution. It is this shift – from human agents making decisions, to human agents and machine agents making decisions together – that illustrates the changing shape of the boundedness of rationality by the agents acting within the organization. Human agents, machine agents, and hybrid teams of human and machine agents each present different contours for the boundedness of rationality. As illustrative examples, human agents are particularly adept decisions that require context and human norms, where machine agents have vastly superior memory storage and number processing capabilities. These present different landscapes of boundedness for the agents making decisions within public organizations.

Furthermore, the boundedness of agent rationality may present a challenge to decision making systems built to leave humans predominantly "out of the loop," or in which multiple specialized AI agents engage in solving joint decision-making tasks. AI processes information, makes decisions, and acts in ways that do not perfectly overlap with their human equivalents: because human observers may struggle to comprehend and control these systems, cooperation may be difficult.

Value misalignment of public organizations

Simon's organizations are information-processing entities, and decisions, based upon assumed or implicit facts and values, are the basic building blocks of this information flow. The co-mingling of facts and values gives rise to decisions, and thus to administrative behavior. As such, an organization should be very concerned with *which facts* and *which values* are brought to bear upon any particular course of action.

Simon's framework highlights the difficulty of aligning and controlling human behavior when organizational goals do not perfectly overlap with those of their human agents, creating problems of authority, motivation, and value alignment. While both Simon and Weber write directly to this challenge, Simon drives home the point that values, as basic input premises, infiltrate at individual decision points throughout organizational flows of action and information processing. These arguments assume that there is some quality of *human decision making* and *human values* that does not naturally arise in the machine agent. Put bluntly, human value systems are contextual and complex, and rooted in subjective human experiences, and thus are unlikely to come forth organically from the manipulation of digital machine processing and code. So, as AI acts as *decision maker* in the

administrative behavior process, these decisions *inherently* follow some value system that is unlikely to match those of the organization or its individual human agents.

The value misalignment of a human–AI collaboration is likely to be a two-way interaction, as either humans input biased training data to the machine, or the algorithms feed biased data to humans. In the presence of AI-supporting tools, the interaction of human bias and algorithm bias may worsen the bias (Sun et al., 2020). Recent empirical studies have found increases in racial and social discrimination in web-searching algorithms (Noble, 2018). Janssen and Kuk (2016) revealed that human faults appear in the workings of algorithms, as, for example, discrimination and inequality.

Additionally, Simon's view on the stimulus and response mechanism of an individual's decision-making behavior explains the uncertainty in the exercise of administrative discretion, which is likely to bias the perceptions of public organizations and their people on emerging technologies, or, even worse, lead them to make the wrong choices. Cognitive biases may confirm their existing perceptions of AI and its value (Huang et al., 2021), similar to what Simon called the mechanisms of behavior persistence, creating internal stimuli for further actions. Furthermore, users may practice automation bias, accepting faulty recommendations based on type II errors (false positives) made by AI and algorithms (Young et al., forthcoming).

Biases may decrease the effectiveness of public performance and increase value misalignment in service provision. Therefore, the introduction of AI agents into organizational information processing and decision-making flow presents unique and important challenges for value alignment throughout a public organization.

The balance of formal and informal communication and authority

Simon reminds us that it is key to use communications to encourage the flow, absorption, and storage of both information and knowledge within organizations. Organizational communications tie closely with organizational levels; formal organizational charts and rules determine the dispersion of hierarchical authority. Formal and informal expressions along the rank of authority play an equally important role in the dispersion of meaningful communication throughout an administrative organization. While formal communications may be decisive, informal authority based around social relationships, organizational power, and other intraorganizational networks enduringly influence how power is conferred and understood and grease the organizational wheels. All formalized systems of authority operating in an environment this complex are, by definition, incomplete tools for dispersing authority required across all decision sets. Similarly, formal information channels play the dominant or decisive role in communication but lack the capacity to manage all needed information flows, motivate personnel, and strengthen team commitment.

The promotion of human–AI collaboration could change the balance of formal and informal communication within public organizations. Considering that the purpose of communication is to reduce information asymmetries, emerging AI tools raise questions about interactions between communicators and the resulting transmission of information for decision making. Here, we propose two possible scenarios. First, the use of a social AI tool, with its personalized predictions, may improve intra-organizational communications and truthful information transmission by spotting individual problems and dissatisfactions

(Schenk, 2021). However, this social algorithm could alternately amplify the curation of certain information and affect trust and communications within organizations (Lazer, 2015). This triad of storage mechanisms (human brains, computer records, and machine learning algorithmic systems), operating in the presence of an artificial intelligent tool with the ability to learn and adjust its solutions based on prior experience and mistakes, broadens the scope of communication problems.

In summary, both Weber and Simon reflected on the control of administrative behavior, considering structure and human psychology as opportunities and challenges for effective, controlled coordination. However, they addressed these questions through differing lenses. Weber emphasized the structures and functions of bureaucratic behavior, whereas Simon dealt with the psychology of information processing and administrative decisions. While Weber used this structured power and control across organized human behavior as a starting point to illustrate the ideal bureaucracy, Simon's logical framework built from the organizational requirement for individual information processing and decision making. Meanwhile, the comparison of these two theoretical views suggests overlapping strategies for managing AI's risks to organizational decision making and effective human control. We explore these strategies below.

Discussion: Dehumanization and loss of control

Weber and Simon highlight distinct and shared challenges to public organizations. In an evolutionary process influenced by rules, humans, and technology, AI is leading the reshaping of resources and communication, and thus decisions and actions. In this trajectory, human judgment and discretion are constricted in many domains. Recent related work suggests that AI may engage in actions that are systematically unaligned with the interests of individual humans or of humanity more generally. Beyond value misalignment, dehumanizing impacts, amplified by AI, are already at play in modern organizations, systematically elevating dangers of administrative evil in public organizations that wield power and direct resources (Young et al., 2021).

While these elements of dehumanization and loss of control in a changing world paint a disturbing picture, the classic theories of Weber and Simon can help us navigate it (Bullock et al., 2022). The first step is to apply Weber's concept of a *Beamte* to the development of bounded, controlled machine agents with certain levels of discretion, similar to the explicit and tacit rules that enable and constrain civil servants. Next, we recommend using Weber's concepts of hierarchy, communication, and specialization as guides for structuring the working tasks of AI or machine agents, and his repulsion for a judging machine to help determine the types of decisions AI are authorized to make.

Simon also gives us tools and guidance for managing administrative organizations as they integrate decision making and action with AI. For Simon, the careful identification of facts and values that guide administrative decision making is a reminder to apply AI with care, and with the knowledge that AI may substitute its own implicit value at any decision point. Simon's notion of complexity and specialization also highlights the types of tasks that AI may usefully address, and paths for controlling its influence. He also highlights the challenges of coworking arrangements and the need for effective communication and

decision making. This analysis suggests that creating effective human and AI teams will be challenging and critical.

It may further be the case that the integration of AI into large public organizations forces us to move beyond the insights of Weber and Simon and reconceptualize larger modern public organizations as their own emergent phenomenon. Bureaucracy scholars have already noted the significant degree of evolution that has occurred among certain types of public organizations, from street-level bureaucracies to systems-level bureaucracies, which predated the fuller integration of AI and machine agents into decision making. Brockman (2020) has described these new large public organizations as "hybrid superintelligences" of human and machine. This combination allows the entity to surpass the intelligence and decision-making capability of any single group of humans.

However they might be known, modern public organizations are clearly evolving with their environment and the tools at their disposal. As AI is dispersed throughout the decision-making functions of these institutions, they may feature greater dehumanization and lower levels of human control. This is, to steal an overused term, unprecedented in the history of public organizations.

Conclusion

Max Weber and Herbert Simon each provided enduring pictures of how organizations decide, act, behave, and learn. Weber produced the classic notion of an ideal bureaucracy as a way to organize human behavior to conduct the business of government. This structure, with its clear lines of hierarchy, authority, and communication, is inhabited by trained professionals who apply their knowledge of the rules and circumstance to the best of their ability.

Simon took the decision itself as the basic premise for organizational action, building out a structure and concepts for hierarchy, authority, and communication. He proposed designing the flow of information throughout an organization so that the most qualified decision maker makes and implements the call. There are, of course, other pictures of public organizations. Prominent among these is Elinor Ostrom's notion of Institutional Analysis and Development, a broader conception of governance that considers organizational interactions in action spaces. While illuminating, this construct is less focused on understanding the structure and decision making of large, bureaucratic organizations.

We have argued that as the external environment has continued to create increasingly capable AI and machines, public organizations have begun to make more systematic use of these technologies, leading to an evolution in many organizations. To predict the direction of this evolution, we turned to Weber and Simon for guidance on how the incorporation of AI into organizational decision-making functions may influence its trajectory. For Weber, the threats were (1) a decreased scope of tasks for human bureaucrats, (2) the loss of human managerial control, and (3) an evolving bureaucracy structure. Using Simon's framework, we argued that AI (1) reshapes the contours of the bounded rationality of agents, (2) presents challenges for value alignment of public organizations with their mission, and (3) changes the balance of formal and informal communication. These evolutionary changes point to

losses in human control and general dehumanization of the decision-making process and corresponding actions.

Furthermore, again based on the classic organizational pictures of Weber and Simon, we learn the importance of restricting the agency of AI through formal hierarchy and communication. AI should only be applied to decisions that would not benefit from human judgment and do not have the potential for further dehumanization and loss of human control. We have also questioned whether these Weber's and Simon's models are as useful as they once were for describing the behavior of modern organizations that conduct the tasks of governments.

With the incorporation of AI and digital networking technologies throughout modern public organizations, it may be better to think of these new entities as hybrids, human and machine entities that are so deeply intertwined as to be inseparable, leading to a decrease in human control and an increase in the dehumanization of decisions and actions. Left unchecked, or steered poorly, these incredibly powerful and large hybrid public organizations may become, forever, misaligned with the interests of humanity and the freedom of individual humans.

References

Agarwal, P. K. (2018). Public administration challenges in the world of AI and bots. *Public Administration Review 78* (6), 917–921. https://doi.org/10.1111/puar.12979.

Arrow, Kenneth J. (1984). The economics of agency. In John W. Pratt and Richard J. Zeckhauser (Eds.), *Principals and agents: The structure of business* (pp. 1–35). Harvard Business School Press.

Bellotti, Marianne. (2021). Helping humans and computers fight together: Military lessons from civilian AI. *War on the Rocks*, March 15. https://warontherocks.com/2021/03/helping-humans-and-computers-fight-together-military-lessons-from-civilian-ai/.

Bertelli, A. M. (2012). *The political economy of public sector governance*. Cambridge University Press.

Bovens, Mark, & Zouridis, Stavros. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review 62* (2), 174–184.

Brockman, John, ed. (2020). *Possible minds: Twenty-five ways of looking at AI*. Penguin.

Bullock, Justin B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration 49* (7), 751–761. https://doi.org/10.1177/0275074019856123.

Bullock, Justin B., Huang, Hsini, & Kim, Kyoung-Cheol (2022). Machine Intelligence, Bureaucracy, and Human Control, *Perspectives on Public Management and Governance*, forthcoming;, gvac006, https://doi.org/10.1093/ppmgov/gvac006

Bullock, Justin B., & Kim, Kyoung-cheol. (2020). Creation of artificial bureaucrats. *Proceedings of European Conference on the Impact of Artificial Intelligence and Robotics* . UK: Academic Conferences and Publishing International Ltd. http://toc.proceedings.com/56779web toc.pdf.

Bullock, Justin, Young, Matthew M., & Wang, Yi-Fan. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity 25* (4), 491–506. https://doi.org/10.3233/IP-200223.

Busch, Peter André, & Eikebrokk, Tom Roar. (2019). Digitizing discretionary practices in public service provision: An empirical study of public service workers' attitudes. 52nd Hawaii International Conference on System Sciences. http://doi.org/10.24251/HICSS.2019.379.

Busch, P. A., & Henriksen, H. Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity 23* (1), 3–28.

Dempsey, Michael. (2021). The navy sub commanded by artificial intelligence. *BBC News*, May 20. https://www.bbc.com/news/business-56993035.

Dunleavy, P. (2005). New public management is dead—long live digital-era governance. *Journal of Public Administration Research and Theory 16* (3), 467–494. https://doi.org/10.1093/jopart/mui057.

Farber, Ken, Lam, Kristine, Taylor, Ellery, Briscoe, Erica, Dretzka, Erica, Scott, John, Sprague, John, Wilkins, Ken, Kyoung-cheol, Eng, Kim Richard, Lee, Steven. (2021). From ethics to operations: Current federal AI policy. Washington, DC: Advanced Technology Academic Research Center. https://atarc.org/wp-content/uploads/2022/01/Current-Federal-AI-Policy-Assessment-FINAL.pdf.

Hambling, David. (2021). U.S. navy destroys target with drone swarm—and sends a message to China. *Forbes*, April 30. https://www.forbes.com/sites/davidhambling/2021/04/30/us-navy-destroys-target-with-drone-swarm---and-sends-a-message-to-china/.

Huang, Hsini, Kim, Kyoung-Cheol (Casey), Young, Matthew M., & Bullock, Justin B. (2021). A matter of perspective: Differential evaluations of artificial intelligence between managers and staff in an experimental simulation. *Asia Pacific Journal of Public Administration 44* (1), 47–65. https://doi.org/10.1080/23276665.2021.1945468.

Huber, J. D., & Shipan, C. R. (2002). *Deliberate discretion? The institutional foundations of bureaucratic autonomy*. Cambridge University Press

Janssen, Marijn, & Kuk, George. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly 33*, 371–377.

Korinek, Anton, & Stiglitz, Joseph E. (2018). Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence: An agenda* (pp. 349–390). Conference Held by National Bureau of Economic Research (2019). University of Chicago Press.

Lazer, David. (2015). The rise of the social algorithm. *Science 348* (6239), 1090–1091. https://doi.org/10.1126/science.aab1422.

Miller, John H., & Page, Scott. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton University Press.

Noble, Safiya Umoja. (2018). *Algorithms of oppression: How search engines reinforce racism*. Illustrated edition. NYU Press.

O'Toole, Laurence J. (1997). Treating networks seriously: Practical and research-based agendas in public administration. *Public Administration Review 57* (1), 45. https://doi.org/10.2307/976691.

Pala, Ali, & Zhuang, Jun. (2019). Information sharing in cybersecurity: A review. *Decision Analysis 16* (3), 157–237. https://doi.org/10.1287/deca.2018.0387.

Putnam, Robert D. (1977). Elite transformation in advanced industrial societies: An empirical assessment of the theory of technocracy. *Comparative Political Studies 10* (3), 383–412. https://doi.org/10.1177/001041407701000305.

Rourke, Francis E. (1972). *Bureaucratic power in national politics*. 2nd ed. Little, Brown.

Sætra, Henrik Skaug. (2020). A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society 62* (August), 101283. https://doi.org/10.1016/j.techsoc.2020.101283.

Schenk, Alicia von. (2021). Predictably satisfied: Contributions of artificial intelligence to intra-organizational communication. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3856479. https://ssrn.com/abstract=3856479.

Simon, Herbert A. (1946). The proverbs of administration." *Classics of public administration*. 7th ed. Wadsworth/Cengage Learning.

Simon, Herbert A. (1979). Rational decision making in business organizations. *The American Economic Review 69* (4), 493–513.

Simon, Herbert A. (1997). *Administrative behavior*. 4th ed. Simon & Schuster.

Simon, Herbert A. (1996). *The sciences of the artificial*. 3rd ed. MIT Press.

Sun, Wenlong, Nasraoui, Olfa, & Shafto, Patrick. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE 15* (8), e0235502. https://doi.org/10.1371/journal.pone.0235502.

Waters, Tony, & Waters, Dagmar, eds. (2015). *Weber's rationalism and modern society*. Palgrave Macmillan US. https://doi.org/10.1057/9781137365866.

Weber, Max. (1922). Bureaucracy. In *Classics of public administration*. 7th ed., 2012, 63–67. Wadsworth/Cengage Learning.

Weber, Max (1978). Economy and Society Vol.2. In Guenther Roth & Claus Wittich (Ed.). Berkeley: University of California Press.

Young, Matthew M., Bullock, Justin B., & Lecy, Jesse D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance 2* (4), 301–313. https://doi.org/10.1093/ppmgov/gvz014.

Young, Matthew M., Himmelreich, Johannes, Bullock, Justin B., & Kim, Kyoung-Cheol. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4* (3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.

CHAPTER 20

AN ECOSYSTEM FRAMEWORK OF AI GOVERNANCE

BERND W. WIRTZ, PAUL F. LANGER, AND JAN C. WEYERER

The Relevance of an Ecosystem Approach to AI and Its Governance

The expectations and fears surrounding artificial intelligence (AI) are discussed more controversially today than almost any other topic. On the part of the information technology (IT) sector in particular, there is enthusiasm for the possibilities of using AI for future products and services (Clifford, 2018). But there is also concern about omnipotent AI: Stephen Hawking even said: "The development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014).

Today the public is getting a first taste of how the AI-based applications can help in our everyday life: the voice assistants from Apple, Amazon, and Google have made it to market maturity and are getting smarter all along; the first cars can drive semi-autonomously; and intelligent software is increasingly being used successfully in the world of business. However, the obvious risks are becoming apparent here. There are the first cases of AI-based discrimination against minorities and the black-box nature of many machine learning approaches makes it difficult to track AI-based decisions and limit the use of private data.

Against this background, the importance of dealing with the topic of AI can hardly be overestimated. The mission is to ensure that the development of AI does not promote the disaster scenarios mentioned at the beginning. The question is whether and how the development and use of AI should be controlled. The all-encompassing scope of AI brings with it particular challenges. It becomes clear that the development of AI should be understood as an ecosystem of its own, in which systematic levels of technologies, actors, and functions are interwoven. In this context, it is therefore important to first gain an understanding of the different aspects of the AI ecosystem and, building on this, to understand what governance can look like. Such requirements therefore require

systems thinking and a corresponding systems governance approach (Burch et al., 2019; Corbett, 2017).

Experts in the public discourse are not only calling for appropriate political governance of those AI ecosystems and for underlying reflection of its governance (Dafoe, 2018). Dwivedi et al. (2019) request a precise governance approach that outlines clear responsibilities and constantly assesses risks involved. Similarly, Gasser and Almeida (2017, p. 60) demand work on the topic of AI governance, assuming that the complex construct of AI "requires new thinking about policy, law and regulation." Floridi (2018, p. 2) states, "Regulating all this has now become the fundamental question." Lastly, the European Parliament Committee on Legal Affairs (2018) calls for developing a guiding framework to deal with current and future developments. So, the question is how the evolution of an AI ecosystem can be managed for the benefit of society. Specifically, it is also about the role of the state, what governance can look like, and who is responsible for it.

To provide a concrete overview, it is important to enter AI in its essence, as it is not a technology per se, but a technology approach that shapes an understanding of technological capability. Following such a view, it is important to distinguish at what stage of development of AI capability governance becomes more important. Thus, while some simple applications of AI do not necessarily cause harm, certain stages of AI development require special attention in the context of governance because of their broader impact on society. There are still few AI frameworks in the literature that claim to provide an integrative overview while keeping the governance perspective in mind. Some frameworks specifically map the need for AI technologies in the private sector, while other approaches look at specific areas of AI-related problems and their governance. With this in mind, the main objective of this chapter is to first develop a comprehensive understanding of an AI ecosystem, of AI governance and, based on this, derive an integrative framework for the governance of AI.

Against this background, the chapter is structured as follows: After the introductory section, we briefly describe the current state of research on AI governance models and frameworks. Subsequently, we derive a common basis for understanding AI ecosystems and its governance. Based on this foundation, we derive a comprehensive framework for AI governance taking into account an ecosystem perspective that links different aspects of AI.

With new developments today, it is not possible to set fixed frameworks, so the policy approach should focus on designing a process that enables meaningful and dynamic governance. A meaningful process is the key approach to reap the benefits of AI and avoid the drawbacks. The framework also encompasses the various AI functions, which are in principle independent of their technological design. A distinction can be made between the sensing, the understanding, and the performative act, which are subject to specific governance concepts. Finally, it is important to consider the technological basis of AI, which is relevant to the governance concept due to aspects such as access to data and the limits of its usability.

In summary, the ecosystem framework presented is meant to provide a comprehensive reference point for the development and implementation of an AI governance strategy and corresponding measures. In addition, an outlook is provided on which aspects of AI governance that are of particular importance for the development path need to be emphasized.

Status Quo AI Governance Frameworks

The study of artificial intelligence has only entered the academic mainstream in recent years with digitization and the actual application of AI approaches (Wirtz et al., 2021). Understandably, the initial focus is on the technical aspects and immediate impacts. Frameworks with a governance focus were therefore less discussed initially but are gaining in importance as society fundamentally needs to adapt to this disruptive technology approach. Consequently, there have been increased efforts to address the topic scientifically (e.g., Gasser & Almeida, 2017; Rahwan, 2018; Scherer, 2016; Winfield & Jirotka, 2018; Wirtz et al., 2020, 2022).

At first glance, the existing governance approaches in the literature have a similar objective, but differ greatly in their concrete areas of application, their scope, and their degree of elaboration (Wirtz et al., 2022). There are some important studies that take a meta-perspective and show which aspects are important in the context of public administration. For instance, Bullock (2019) clarifies that, in principle, an improvement in intelligence discretion and decision-making can also be expected to improve the overall quality of administration. However, the author argues that it is important to examine the organizational and legal context and, in particular, the distribution and characteristics of tasks in terms of complexity and uncertainty, as these will determine the extent to which the use of AI makes sense. In this context, Young et al. (2019) provide a theoretical framework to identify what impact AI has on administration. They distinguish artificial discretion with human discretion in different tasks and environmental complexity, highlighting benefits and concerns at the same time. Furthermore, Young et al. (2021) present an ethical framework to analyze the impact of AI in public organizations and to assist in decision-making regarding AI use in the public sector. They use existing theoretical challenges such as agency theory; information problems in the form of adverse selection and moral hazard; and different levels of analysis, such as individual (micro), organizational (meso), and cultural (macro), to make fundamental propositions about the use of AI in the public sector. Finally, the work of Wirtz et al. (2020, 2022) has attempted to conceptually consolidate prior AI governance research in the context of public administration in the form of challenge or risk-oriented integrated AI governance frameworks.

Overall, AI governance research reveals a variety of substantial governance challenges and points to the need for a differentiated elaboration of an adequate public AI policy to guide the successful commissioning of an AI-supported computing system, thereby ensuring public value creation through AI and protecting the public from a malfunctioning supercomputer.

Towards an Understanding of AI Ecosystems and Its Governance

The increasing complexity of man-made interaction systems, such as the internet, has contributed to structures being repeatedly referred to by the term ecosystem. The term

refers to a concept inspired by biology and ecology (Scaringella & Radziwon, 2018) describing "the whole system . . . including not only the organism-complex, but also the whole complex of physical factors forming what we call . . . the habitat factors in the wildest sense" (Tansley, 1935, p. 299). Moore takes this understanding and applies it to firms in an economic system, outlining how firms "coevolve capabilities around a new innovation: they work cooperatively and competitively to support new products, satisfy . . . needs, and eventually incorporate the next round of innovations" (Moore, 1993, p. 76).

In contrast to the value chain view, an economic ecosystem understanding includes all process-influencing actors, including those outside the linear value creation process (Clarysse et al., 2014; Iansiti & Levien, 2004). Value creation is no longer characterized by linearity (Moore, 1996) but arises in a network of interconnected actors and organizations that use their specific capacities and competences to jointly realize value creation (Christensen & Rosenbloom, 1995). The association in an ecosystem thus enables value creation that none of the actors involved would achieve individually (Adner, 2006). However, the planning and control of the ecosystem is not defined in this context.

If we apply this perspective to AI, the following arguments are important:

(1) Individual and specific AI systems are comparatively easy to control and limit in their risk potential. If one understands ecosystems as proprietary platform environments with a high degree of compatibility that attracts a multitude of actors who build their activities on the platform environment (Apple, Microsoft, etc.), then it must follow that AI must first be responsibly managed by the owners. This brings this area into the realm of private sector AI governance. Companies that use AI applications must be held accountable for their outputs and results.
(2) Another understanding refers to a macro perspective that includes different actors, systems, and processes that integrate their AI activities. Such an understanding refers to a networked context of multiple actors that form a dynamic, complex, and relational structure representing all transactions that are more effective together than individually. The challenge is the network of different AI systems whose results are further used by other AI systems. This ecosystem perspective creates a need for a broader and public AI governance perspective.

In order to derive a basic understanding of AI and, in particular, with regard to the development of a concept for the governance of AI, it is important to recognize and understand its complexity. AI is often understood as a cross-cutting technology or technology approach, and it can be differentiated according to various criteria. Most definitions of AI refer to an understanding of artificial systems that mimic human-like intelligence independent of human assistance: "AI refers to the capability of a computer system to show human-like intelligent behavior characterized by certain core competencies, including perception, understanding, action, and learning" (Wirtz et al., 2019, p. 599). This understanding involves independent thinking and learning and, specifically, abstract problem-solving abilities. The relevant core intelligence defining competencies derived from its human origin can thus be divided into the components of sensing, comprehension and learning, and action, whereby an underlying assumption always refers to the artificiality and thus abstraction and learning independence of human interference (Weyerer & Langer, 2019; Wirtz & Müller, 2019).

The sensing aspect and the action aspect are thus an integral part of the comprehension and learning component of artificial intelligence, but are not necessarily technically demanding because sensing can also refer to simple data sets (such as texts, sensor measurements, etc.), and deciding or acting is merely a competence extension of the machine output. The area of comprehension or learning, on the other hand, is technically challenging and is often considered separately with the term machine learning. Machine learning is thus not just a conceptual distinction, but an area of AI that refers to automated, repetitive learning, often based on natural language processing (NLP) (Cambria & White, 2014).

All three areas can be considered separately in terms of governance. In the area of sensing, there is always the question of the extent to which data is biased and whether sensitive data can be made available for use.

At the core of AI is the ability to learn autonomously. In the case of machine learning, the biggest challenge is that modern approaches (neural networks, deep learning) produce meaningful learning results, but their learning process is not comprehensible. Against this background, it is often referred to as a black box. Buiten (2019), points out that research should be conducted in particular on the "black box" character of AI and that the emergence, prevention, and regulation of the associated risks are important. Thus, this opacity is a problem that must be addressed with appropriate governance to ensure traceability, reproducibility, acceptance, and last but not least, reliability.

Finally, when it comes to independent decision-making and action, there is the big question of responsibility, the social and legal basis, and, of course, human control. Figure 20.1 distinguishes between the areas of AI sensing, AI comprehension/learning, and AI action, and shows the respective aspects for corresponding governance.

Besides the different parts of AI (sensing, comprehension/learning, and action), another important differentiation refers to the development stage of AI. Here, the independence of technology from human action is the basis for differentiation. In general, it can be assumed that the more autonomous the technology acts and the more independent it is, the more unspecific its scope of application (Kaplan & Haenlein, 2019).

Considering all those different aspects of AI and their governance implications, it is important to consider also the governance motive. From a liberal understanding, governance in the sense of promoting or regulating AI cannot be an end in itself, but must be oriented exclusively towards concrete advantages or dangers that result from the development and application of AI. In this context, it is not only the AI itself that is relevant, but also the environment that is substantially changed by the AI to its advantage or harm. In principle, AI governance therefore must go beyond the mitigation of risks, as considered by Jelinek et

Artificial Intelligence		
AI Sensing	AI Comprehension / Learning	AI Action
AI Governance Aspects: Access to data, non-biased data sources, privacy, ...	**AI Governance Aspects:** Traceability, reproducibility transparency of the black box, reliability, ...	**AI Governance Aspects:** Legality, responsibility, human control, ...

FIGURE 20.1 AI governance for the areas of sensing, comprehension/learning, and action (Bataller & Harris, 2016; Wirtz & Müller, 2019)

al. (2021, p. 142), who see AI governance as a "widely shared, international approach" that exclusively aims to prevent and mitigate AI risks. Against this background, we propose a broader understanding of AI governance that takes into account the opportunities of AI ecosystems for society and sets appropriate incentives for their use. Furthermore, it is important to consider the complex organizational, technological, legal, and ethical forms of governance as equally relevant.

Definition of AI governance

> *AI governance is the management of AI development and application by setting rules and frameworks for action, taking into account the interaction effects of the potential evolution towards an AI ecosystem. It addresses real and potential impacts of AI on society as well as on institutional and normative frameworks for government and business by defining rules for technological, informational, political, legal, social, economic, and ethical action. Depending on the field of application and the expected risks and benefits, different means are chosen to restrict as well as promote the use of AI.*

In the context of AI governance, it is also important against the backdrop of a liberal order of values that the application of regulation takes place in a balanced manner (Jobin et al., 2019). A distinction can be made between binding and non-binding rules or soft and hard law, which can be chosen according to the risk potential. Governance in a dynamic AI ecosystem relies on both approaches. Depending on the challenge, solutions must be found between fixed and binding laws to voluntary standards, norms, guidelines, or soft law (Marchant, 2019; Villasenor, 2020).

Moreover, the perspective of governance remains. The AI governance definition presented refers to society being the primacy of governance. However, there is also an increasing use of AI in regulation and governance (Coglianese, 2018; Coglianese & Lehr, 2016). AI governance can therefore also be understood as AI governing existing systems and people—this understanding is not further adhered to in this chapter.

An Ecosystem Framework of AI Governance

An AI governance ecosystem involves certain ecosystem members whose roles, capabilities, interests, and actions are specific to AI governance and who all affect the process of governance. AI neither affects a single actor nor is governed by one, but rather involves a variety of actors. Government, industry, civil society, academia, etc. all play a certain role in AI governance on a national and global level. Together they form a network of actors who have connected interests and cooperate and compete with each other in order to survive, thus resembling a biological ecosystem "that gradually moves from a random collection of elements to a more structured community" (Moore, 1993, p. 76). The merging of the actors and their capabilities in an ecosystem lead to outcomes or value creation that none of

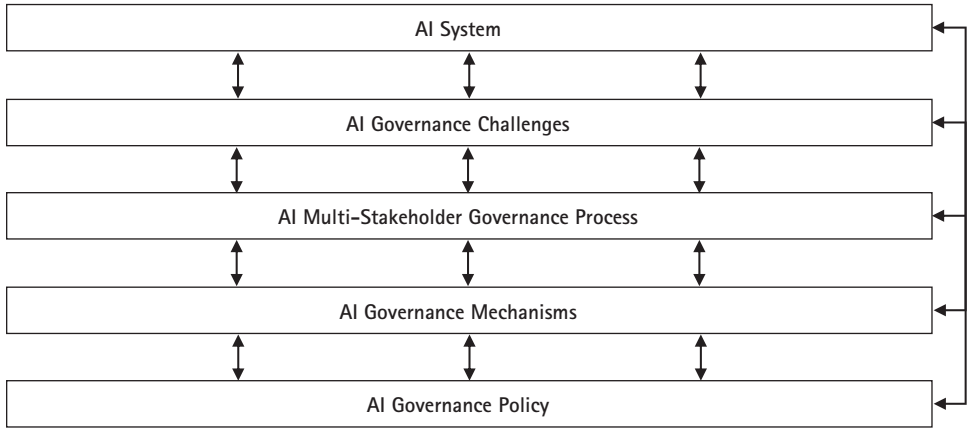

FIGURE 20.2 General framework structure and layers

them would have achieved individually (Adner, 2006). The main features of AI ecosystem governance can be described in terms of five levels: (1) AI system, (2) AI challenges, (3) AI multi-stakeholder governance process, (4) AI governance mechanisms, and (5) AI governance policy. The order presented is for illustrative purposes only and does not represent a strict order. An ecosystem is to be understood as an interconnected complex system, which is why the different levels also interact with each other. For example, the AI system can directly influence the AI governance policy through its technological development and vice versa, as governance can define restrictions and incentives that in turn influence the AI system. Each level thus has a dynamic influence on all others, which is why a meta-perspective that does justice to the complexity is so important.

Figure 20.2 illustrates the general structure and the layers of the ecosystem framework of AI governance.

The AI system is the object of governance and constitutes the starting point of the framework (AI system layer). The operation of an AI system in the field involves a variety of common challenges that reflect the concrete aspects in need of governance. They determine the specific governance requirements and fields of action in AI governance (AI governance challenges layer). These AI governance challenges trigger a governance process that is subject to multi-stakeholder input and dialog, and in which the aspects of governance are framed, evaluated, and prepared for AI policy-making (AI multi-stakeholder governance process layer). Decision makers and policymakers may draw on different governance mechanisms in order to develop governance measures (AI governance mechanisms layer). The resulting governance measures address certain AI governance challenges and their respective field of action, forming in their entirety an AI rulebook (AI governance policy layer).

AI system

The AI system layer entails the AI system in terms of an AI application or service as well as its underlying technology infrastructure. An AI system is composed of the three

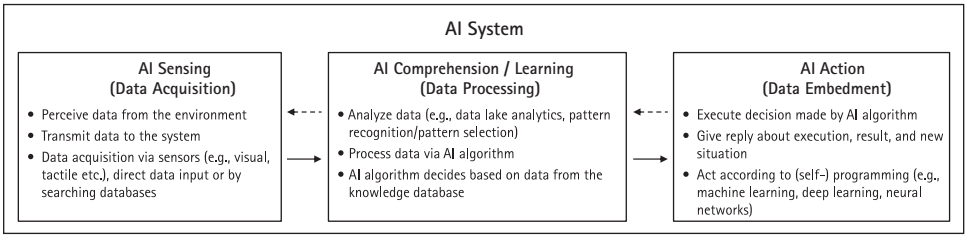
AI System
AI Sensing
(Data Acquisition)
• Perceive data from the environment
• Transmit data to the system
• Data acquisition via sensors (e.g., visual, tactile etc.), direct data input or by searching databases
AI Comprehension / Learning
(Data Processing)
• Analyze data (e.g., data lake analytics, pattern recognition/pattern selection)
• Process data via AI algorithm
• AI algorithm decides based on data from the knowledge database
AI Action
(Data Embedment)
• Execute decision made by AI algorithm
• Give reply about execution, result, and new situation
• Act according to (self-) programming (e.g., machine learning, deep learning, neural networks)

FIGURE 20.3 AI system layer (Bataller & Harris, 2016; Wirtz & Müller, 2019)

above-mentioned core capabilities of sensing, comprehension/learning, and action (Bataller & Harris, 2016). These capabilities manifest themselves in the functional processes that take place within an AI system and also determine the functional logic of the underlying technology infrastructure. Figure 20.3 depicts the AI system layer and its three core capabilities.

Sensing refers to the process of perceiving the environment with its objects and subjects through sensory inputs, as well as processing this information from the environment (Bataller & Harris, 2016; Russell & Norvig, 2016). In fact, few AI systems today have direct sensory input. Most AI systems today work with existing data sets. However, many data sets rely on information from sensors, such as the keyboard and mouse, to sense human input. Other data is often based on data sets from sensory systems in the form of sensing devices (e.g., cameras, microphones, touch sensors, etc.) and software (e.g., computer vision and audio processing) that can capture and process environmental information in the form of data and store this data in knowledge bases for further processing and use. For example, self-service kiosks for border control are equipped with intelligent facial recognition technology to match passengers with their passport information and the data collected by the government on each passenger entering the country. Besides this primary data acquisition by means of its sensory system, an AI system can also perceive information via a secondary data acquisition, namely through human data input or by retrieving data from already existing databases (Wirtz & Müller, 2019). In a broader sense, sensing may therefore refer to all activities and processes of data acquisition through the AI system (Wirtz et al., 2020).

The second core capability or step of the functional process of an AI system refers to the comprehension and learning of the acquired information and thus the processing of the respective data stored in the knowledge databases. To understand the new data and learn from it, the data needs to be further processed and analyzed by means of applicable technologies and software solutions (e.g., inference engines, natural language processing, data analytics, and knowledge representation). Such technology solutions are used, for example, in connection with AI-based smart speakers (e.g., Amazon Echo). To process and analyze the data and to build a knowledge representation, the learning algorithm of the AI system structures the new data and integrates it with already existing data in a common knowledge base. The knowledge base is then analyzed for potential patterns in order to draw conclusions and deduce a possible decision or course of action.

The knowledge base and the embedment of data into the greater functional work routine of the AI system play a vital role for the behavior of the AI system, making it capable of deciding and acting. The AI system uses the data embedded, or in other words, the insight gained within the former stage of comprehension and learning, to make a decision and to perform an action. It can be distinguished between two main forms of action. On the one

hand, the AI system can use information as internal output and adapt its learning algorithm accordingly. On the other hand, the AI system can deliver external output in terms of reporting the insights gained from the data to the user or accomplish any other task for which it was designed, such as steering a car or answering a customer query (Wirtz et al., 2020; Wirtz & Müller, 2019).

As shown in the beginning, it is essential to distinguish between these three core capabilities, or functional AI processes, for the sake of effective governance because each is associated with certain aspects of governance. As a cross-cutting technology, putting AI into practice may lead to a variety of challenges that may cause harm to humans, other living beings and the environment.

AI governance challenges

There is a plethora of potential governance challenges of AI mentioned in the literature (Dwivedi et al., 2019; Engstrom et al., 2020; Wirtz et al., 2019). The aim of this layer is to outline the major and most pressing governance challenges that can be expected from the introduction of AI in different domains. For this purpose, the AI governance challenges are separated into different clusters and thus made comprehensible. This classification makes clear the broad sphere of AI governance and its fields of action. According to Wirtz et al. (2022), AI governance challenges can be classified into six main categories: (1) technological and data, (2), informational and communicational, (3) economic, (4) social, (5) ethical, and (6) legal and regulatory. Figure 20.4 depicts the areas of AI governance challenges and summarizes the above-mentioned aspects.

Technological and data

Technological and data-related governance challenges refer in particular to the loss of control of AI systems and ensuring the functional safety and favorable impact of an AI system, which requires it to be resilient against negative environmental influences and human manipulation (Bostrom & Yudkowsky, 2014). Two main challenges and sources of malfunctioning and adverse effects of an AI system are programming errors and data-related issues. Programming errors may result from a lack of AI expertise that is a rare and

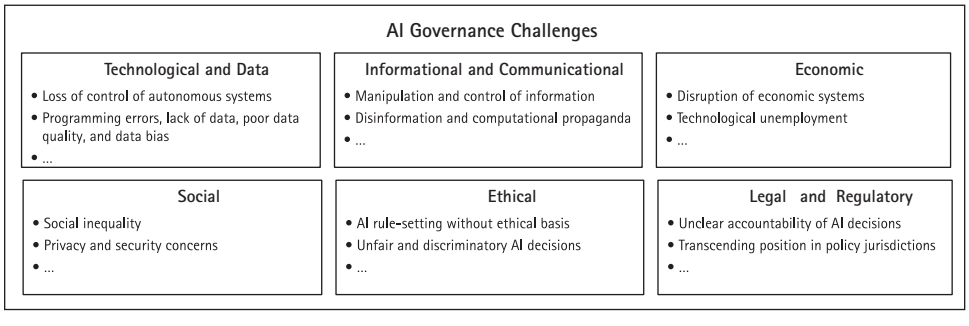

FIGURE 20.4 AI governance challenges layer (Wirtz et al., 2022)

highly demanded capability in the market. Data-related issues especially refer to a lack of data and poor data quality (Dwivedi et al., 2019). If these fundamental input factors of the AI system are not appropriate, the goodness of an AI system is at risk, true to the motto of "garbage in, garbage out."

Informational and communicational

Informational and communicational governance challenges mainly refer to the control and manipulation of information provision through AI systems. News and media recommendation algorithms amplify and suppress certain content determining what the audience can see and cannot see, which "might create a monoculture, in which users get trapped in their 'filter bubble' or 'echo chambers' " (Bozdag, 2013, p. 209). This also opens the way for algorithmic disinformation campaigns or computational propaganda, in which social bots are used to spread fake news on social media or online platforms in order to manipulate public opinion (Coglianese & Lampmann, 2021; Coglianese & Lehr, 2019; Howard et al., 2018; Weyerer & Langer, 2019; Wiesenberg & Tench, 2020).

Economic

Economic governance challenges may occur on macro- and microeconomic levels. The major macroeconomic challenge refers to the potential disruption of economic systems as a consequence of the increasing introduction of AI systems and automation in the economy (Boyd & Wilson, 2017). In particular, human workforce replacement and technological unemployment through AI may lead to a disruption of the labor market, but may also spill over to other economic domains (e.g., taxation) (Gasser & Almeida, 2017; Wright & Schultz, 2018). Mass unemployment means losing taxpayers and a negative impact on the economic system as a whole. Microeconomic challenges refer to organizational risks. Human replacement by AI may lead to a loss of knowledge and control over business processes, which in turn may negatively affect organizational performance. Other important organizational challenges include the lack of AI skills and strategy, the affordability of AI-related costs, and organizational resistance against AI-related issues, such as data sharing (Dwivedi et al., 2019).

Social

A major social governance challenge arises from the clash between technological unemployment and the economic success of AI. Mass unemployment, on the one hand, and increased profits, on the other, may fuel social inequality expanding the gap between rich and poor (Boyd & Wilson, 2017). Privacy and security concerns in society are also of great importance. AI systems are vulnerable to cyber-attacks and illicit surveillance, while increasingly moving closer to the individual and becoming part of their private living environment (e.g., smart speaker, wearables) (Calo, 2010, 2012). The resulting privacy concerns also undermine social acceptance of and trust in AI systems among citizens, which are vital for the success of AI and thus also constitute a major challenge of AI governance (Wirtz et al., 2019).

Ethical

Ethical challenges of AI have received great attention in the literature. The main issue here is that AI systems make decisions and rules for humans without an ethical basis. This is particularly relevant in life-or-death situations that AI systems may face (e.g., automated driving, autonomous weapons). Embedding ethical principles into AI systems and ensuring that they decide and act according to human ethical values remains a major governance challenge (Wirtz et al., 2019). This also includes that AI systems do not violate the ethical principles of fairness and equality (Thierer et al., 2017). Yet, unfair and discriminatory behavior of AI systems against minorities has increasingly made the headlines in recent years (Weyerer & Langer, 2020), ranging from racial and gender bias in predictive policing and online search outputs to discriminatory decisions against disabled patients in healthcare. The reason for this is that the AI system adopts biased input from its environment, for example, human prejudices reflected in the programming of the AI system and its underlying data basis, which then become part of its decisions and actions (Weyerer & Langer, 2019). Another aspect is the automated measurement of human competence and capacity, and the resulting social and reputational scoring approaches, which raises ethical challenges (Langer, 2020).

Legal and regulatory

The increasing opacity and autonomy of the learning and decision-making process of AI systems makes it impossible for developers and operators of an AI system to control and predict all of its decisions and actions (Wirtz et al., 2020). This raises the challenge of determining who is to be held legally accountable for the outcomes of the AI system and their consequences, in particular in case of failure (Reed, 2018; Scherer, 2016).

Further regulatory complexity is added due to the intertwining of AI with different technologies putting AI governance into a transcending position in policy jurisdictions. This complexity of the regulatory AI landscape and the scattered responsibilities constitute a great challenge in achieving a concerted and effective governance approach in the current AI governance vacuum. In this connection, it is also challenging to live up to the full scope of AI governance and to not oversee relevant governance aspects (Gasser & Almeida, 2017; Thierer et al., 2017). Developing and implementing appropriate governance measures in society that address these fields of actions and their respective challenges requires a governance process, in which the different stakeholders engage in a policy-making discourse.

AI multi-stakeholder governance process

The AI governance process can be understood as an element that links the challenges to appropriate governance measures. In the case of new developments, it is impossible to set fixed frameworks today, so the approach should focus on designing a process that enables meaningful and dynamic governance. A meaningful process is the most important approach to reap the benefits of AI and prevent the drawbacks. This means the process has to account for the complexity and dynamism of the AI landscape. It therefore has to be

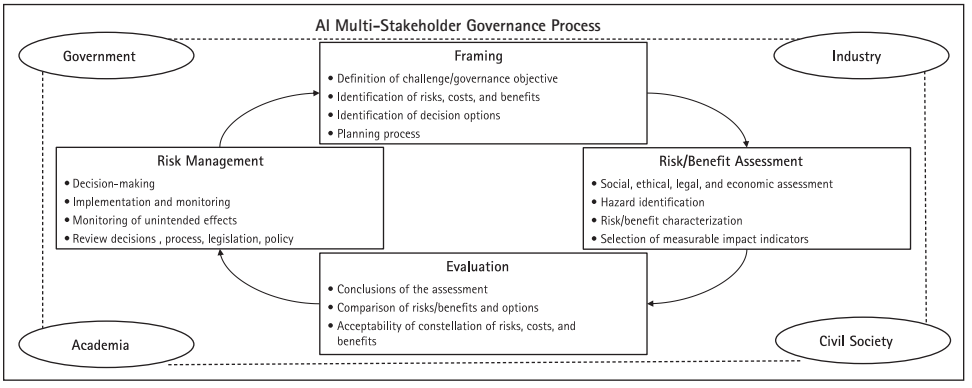

FIGURE 20.5 AI multi-stakeholder governance process (Wirtz et al., 2020)

collaborative and include various stakeholders, in particular, government, industry, academia, and the civil society (Linkov et al., 2018). Government alone cannot keep up with the progress of AI (commonly referred to as "pacing problem") and cannot regulate AI in an effective and comprehensive manner (Guihot et al., 2017; Wallach & Marchant, 2018). In addition, the government process needs to allow for continuous adaptation in order to respond to change and new developments. This can be achieved through a recursive process design including the stages of (1) framing, (2) risk/benefit assessment, (3) evaluation, and (4) risk management (Wirtz et al., 2020). Figure 20.5 illustrates this AI multi-stakeholder governance process.

In the framing stage, the various stakeholders convene to elaborate and define a common understanding of the challenge. This involves a common problem definition in which risk, costs, benefits, and decision options are identified, as well as the formulation of the governance objective they seek to achieve. The framing stage also involves the identification of decision options and determines the further planning process. In the next stage, the risks and benefits are systematically and comprehensively assessed in terms of their social, ethical, legal, and economic impact. This not only requires a clear definition of risks and benefits but also the selection of appropriate and measurable impact indicators. Once these individual assessments of risks and benefits are complete, the evaluation stage begins. Here, the risks and benefits, as well as potential courses of action, are subject to a comparative analysis and are weighed up against each other. The insights and implications from this evaluation constitute the basis of governance decision-making and action, which is shaped and determined in the last process stage of challenge management. Here, the evaluation outcomes are carefully reviewed and a consensual course of action, including the selection of appropriate governance mechanisms, is agreed upon, ultimately leading to the introduction of governance measures.

The governance core aspects (challenges, measures, process, etc.) are subject to continuous monitoring. This is necessary in order to detect potential problems and need for change, which then may induce a reframing of the governance situation and thus a restart of the governance process. To reach a governance outcome in terms of an AI governance policy, decision makers and policymakers can use different governance mechanisms, which are described in the following layer.

AI governance mechanisms

Governance mechanisms can be understood as instruments through which governance is manifested and comes into force. According to Wirtz and Müller (2022) three types of mechanisms can be distinguished: (1) contractual, (2) relational, and (3) technological. Figure 20.6 illustrates the types of AI governance mechanisms.

Contractual governance mechanisms refer to enforceable formal contracts in which the rights, obligations, and roles of the contracting partners are clearly defined in order to prevent and alleviate misunderstandings with regard to activities and procedures (Vandaele et al., 2007). Accordingly, contractual governance mechanisms include legal standards and contracts and formalized guidelines and procedures.

Relational governance mechanisms are rather informal in nature and refer to building good relationships among exchange partners, as well as guiding their exchanges through trust and social norms (Poppo & Zenger, 2002). This comprises collaborative or functional relations, informal exchanges, and agreements.

Besides these more traditional mechanisms of governance that are well-established in the literature, a third type arises in the form of technological governance mechanisms that are particularly relevant in the context of AI. Lumineau et al. (2021, p. 506 have specified technological governance mechanisms in the context of blockchain technology, describing them as "a self-contained and autonomous system of formal rules . . . automatically enforced by the underlying . . . network." Similarly, technological governance mechanisms may be integrated into AI systems resembling an autonomous scheme of formal, code-based rules that govern activities and interactions. Decision makers and policymakers can use these governance mechanisms in order to develop governance measures and a corresponding AI governance policy.

AI governance policy

The AI governance policy can be understood as the body of rules that governs all AI systems and AI-related activities of the actors in the ecosystem. As such, the policy integrates all governance measures or remedies for the challenges posed by AI. Because remedies and challenges can be seen as counterparts to each other, it appears reasonable to organize them in the same way. According to Wirtz et al. (2022), it can be distinguished between (1) technological and data, (2), informational and communicational, (3) economic, (4) social, (5) ethical, and (6) legal and regulatory remedies. These categories can be viewed as governance areas or fields of action that together constitute the scope of the AI governance policy. Figure 20.7 shows the AI governance policy layer and its different fields of action.

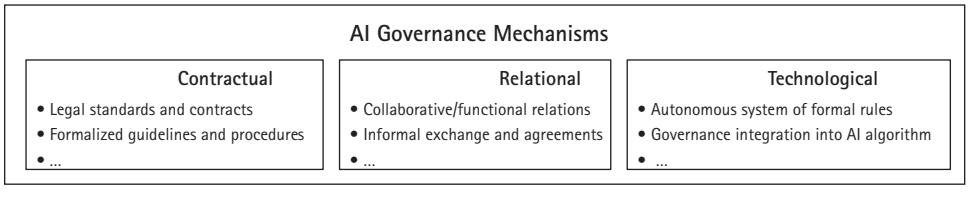

FIGURE 20.6 AI governance mechanisms (Wirtz & Müller, 2022)

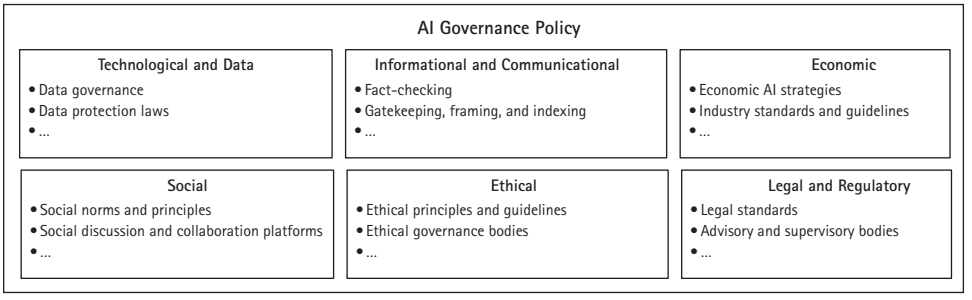

FIGURE 20.7 AI governance policy (Wirtz et al., 2022)

To convey a better understanding of how an AI governance policy may look like, we describe exemplary governance measures in each field of action in the following.

Technological and data

An important governance measure in this area refers to the concept of data governance (Janssen et al., 2020). Data governance is an organizational approach to determine the decision rights and responsibilities with regard to data management in organizations. It formalizes and enforces data management policies and processes to ensure and enhance data quality, involving both contractual (e.g., data sharing agreements) and relational (e.g., training programs) elements (Abraham et al., 2019). Another vital aspect in particular for addressing the privacy challenge of AI are data protection laws, such as the General Data Protection Regulation (GDPR) introduced in the European Union in 2018.

Informational and communicational

A major governance approach in this context refers to fact-checking, which aims at verifying publicly provided information. It is usually carried out by independent NGO's but also by media platforms (e.g., Facebook). Fact-checking may not only be performed by humans but also by AI systems themselves, thus also being exemplary for technological governance mechanisms (Lewandowsky et al., 2017). For example, Facebook deletes fake news regarding the coronavirus pandemic by means of AI, showing warnings in connection with respective posts. Fact-checking may act as a deterrent to actors engaging in AI-related disinformation campaigns (Nyhan & Reifler, 2015) and may contribute to breaking up filter bubbles and echo chambers (Rehm, 2018). Furthermore, adapting traditional media regulation measures, such as gatekeeping, framing, or indexing may also be effective in containing manipulation and control of information through AI bots (Bennett & Pfetsch, 2018).

Economic

A key factor of AI governance is strategy. Governments and organizations need clear and comprehensive strategies and objectives in terms of an AI agenda when it comes to coping

with economic AI challenges. A number of governments across the globe have implemented international and national AI strategies and initiatives in recent years to guide the development and deployment of AI towards economic growth and beneficial social impact (Future of Life Institute, 2021). This involves, for instance, investment plans to promote AI R&D activities, education and computing infrastructure, as well as recommendations and strategies on how to handle AI-related workforce challenges.

In addition, organizations need an AI change management strategy that prepares and accompanies their transition to AI (e.g., workforce training, transparency measures etc.) to meet challenges, such as lack of AI skills or resistance of employees and customers against AI. Further important economic governance measures are in particular so-called soft-law measures including industry standards, certification programs, best practices and codes of conduct (Wallach & Marchant, 2019).

Social

Social AI governance measures refer to creating social institutions such as social norms and principles that guide AI development, deployment, and legislation. Japan, for instance, has developed social principles of human-centric AI, including principles such as privacy, security, fairness, accountability, and transparency (Mitchell et al., 2020). Creating collaborative platforms and public participation formats (e.g., workshops, surveys) that bring together key stakeholders from society and synthesize a broad spectrum of relevant perspectives is vital for the development of such social governance measures. The Partnership on AI (PAI), for instance, seeks to provide such an open collaborative platform to promote best practices and advance the understanding of AI in society.

Ethical

Similarly, a common ethical governance measure pertains to the formulation and implementation of ethical guidelines and principles to guide AI action. Ethical guidelines have been proposed by various actors from government, industry, and academia. For example, the European Commission has set up an expert group that developed the "Ethics Guidelines for Trustworthy AI" (European Commission, 2019); the Future of Life Institute provides the "Asilomar AI Principles," which include different ethics and values (Future of Life Institute, 2017); and Google and Microsoft have also committed themselves to ethical principles with regard to their AI activities (Google, 2021; Microsoft, 2021). Ethical governance bodies in the form of steering committees or ethics commissions that elaborate ethical guidelines and monitor compliance in organizations are essential in this context (Wirtz & Müller, 2019).

Legal and regulatory

Legal governance measures refer to any legal standard or legally binding regulation (e.g., laws, directives) regarding AI. Given the "wait and see" approach to AI regulation of most governments, pertinent AI laws and regulations are scarce, while other aspects related to AI, in particular, data protection and autonomous vehicles, have received more legislative attention (Walch, 2020).

However, more targeted AI legislation is slowly picking up pace. For example, The National Artificial Intelligence Initiative Act of 2020 (NAIIA) is the first legislative manifestation of the national AI strategy of the U.S., aimed at keeping the country at the cutting edge of AI research and development. In addition, The Algorithmic Accountability Act, initially introduced in 2019, has been reintroduced in 2022 in the U.S. to fight algorithmic bias and discrimination. This pending federal legislative approach is the first that seeks to govern AI systems in general and not a specific technological area (e.g., data, autonomous vehicles). The European Union is moving more quickly toward comprehensive AI legislation and has recently set a landmark in global AI regulation efforts by introducing a proposal for the first legal AI regulation framework (European Commission, 2021).

Furthermore, an important element of governance refers to establishing government authorities that are in charge of regulating transparency, quality, and safety issues (Doneda & Almeida, 2016). The U.S. government, for instance, launched the National Artificial Intelligence Initiative Office as a new agency under the NAIIA to oversee and implement the national AI strategy. The legal and regulatory governance measures constitute the closing point of an AI governance policy. Figure 20.8 summarizes the above-mentioned layers and aspects of governance, depicting an integrated ecosystem framework of AI governance.

Outlook

This chapter provides a systemic understanding of a complex and dynamic AI ecosystem and its governance. Although many areas interact in unmanageable ways, it is important to identify structural components. Similar to the biological ecosystem, "[a]ctually the systems we isolate mentally are not only included as parts of larger ones, but they also overlap, interlock and interact with one another. The isolation is partly artificial, but is the only possible way in which we can proceed" (Tansley, 1935, p. 300).

The importance of this structural identification lies in the fact that it allows us to check whether changes in one component go along with changes in other components. In a legislative process, for example, it is possible to check systematically what effects a law could have on the five components and whether the latter trigger further feedback effects. The governance of complex and dynamic ecosystems thus requires a systemic perspective.

The five basic or complementary components are themselves partly immature and require further regulatory development. This intertwining of AI with other enabling technologies and governance sub-areas places AI governance in an overarching position of political responsibility. Coordinating and bringing together the respective efforts into a concerted, effective global multi-stakeholder approach will remain a major challenge in the future. Structuring and systematizing the AI governance ecosystem and its core elements into a common framework is essential for a better understanding and represents a first step towards this difficult task.

The rapid development and increasing use of AI in recent years has brought with it a number of opportunities and challenges for society that draw attention to the need for strong and comprehensive governance. As depicted, an AI ecosystem can be described as a melting pot of diverse technologies, applications, and actors that defines the eclectic, disruptive, and cross-cutting nature of AI. This suggests that AI governance, which aims to shape that

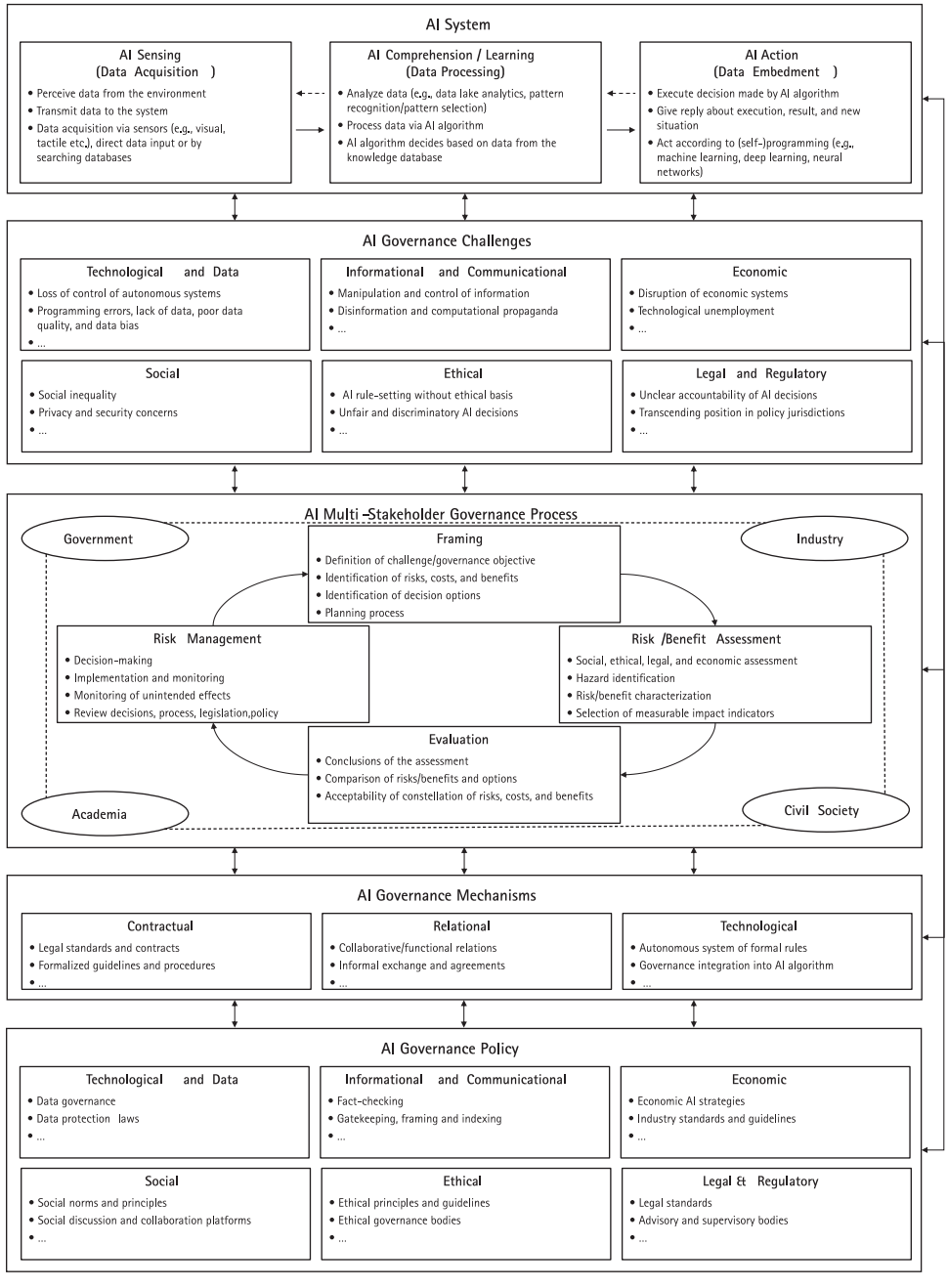
AI System
AI Sensing (Data Acquisition)
• Perceive data from the environment
• Transmit data to the system
• Data acquisition via sensors (e.g., visual, tactile etc.), direct data input or by searching databases
AI Comprehension / Learning (Data Processing)
• Analyze data (e.g., data lake analytics, pattern recognition/pattern selection)
• Process data via AI algorithm
• AI algorithm decides based on data from the knowledge database
AI Action (Data Embedment)
• Execute decision made by AI algorithm
• Give reply about execution, result, and new situation
• Act according to (self-)programming (e.g., machine learning, deep learning, neural networks)
AI Governance Challenges
Technological and Data
• Loss of control of autonomous systems
• Programming errors, lack of data, poor data quality, and data bias
• ...
Informational and Communicational
• Manipulation and control of information
• Disinformation and computational propaganda
• ...
Economic
• Disruption of economic systems
• Technological unemployment
• ...
Social
• Social inequality
• Privacy and security concerns
• ...
Ethical
• AI rule-setting without ethical basis
• Unfair and discriminatory AI decisions
• ...
Legal and Regulatory
• Unclear accountability of AI decisions
• Transcending position in policy jurisdictions
• ...
AI Multi-Stakeholder Governance Process
Government
Framing
• Definition of challenge/governance objective
• Identification of risks, costs, and benefits
• Identification of decision options
• Planning process
Industry
Risk Management
• Decision-making
• Implementation and monitoring
• Monitoring of unintended effects
• Review decisions, process, legislation,policy
Risk /Benefit Assessment
• Social, ethical, legal, and economic assessment
• Hazard identification
• Risk/benefit characterization
• Selection of measurable impact indicators
Evaluation
• Conclusions of the assessment
• Comparison of risks/benefits and options
• Acceptability of constellation of risks, costs, and benefits
Academia
Civil Society
AI Governance Mechanisms
Contractual
• Legal standards and contracts
• Formalized guidelines and procedures
• ...
Relational
• Collaborative/functional relations
• Informal exchange and agreements
• ...
Technological
• Autonomous system of formal rules
• Governance integration into AI algorithm
• ...
AI Governance Policy
Technological and Data
• Data governance
• Data protection laws
• ...
Informational and Communicational
• Fact-checking
• Gatekeeping, framing and indexing
• ...
Economic
• Economic AI strategies
• Industry standards and guidelines
• ...
Social
• Social norms and principles
• Social discussion and collaboration platforms
• ...
Ethical
• Ethical principles and guidelines
• Ethical governance bodies
• ...
Legal & Regulatory
• Legal standards
• Advisory and supervisory bodies
• ...

FIGURE 20.8 An ecosystem framework of AI governance

ecosystem for the benefit of society, must itself take multi-layered and complex approaches. This chapter shows which issues are important without implying a linear process.

The speed of AI progress—especially when there are interactions between AI systems—can be faster than existing governance and regulation (see "pacing problem," Marchant,

2011; Marchant et al., 2013). This shows that AI is more than a simple game changer. It can be assumed that the introduction of AI represents a disruption of existing development phases and can only be compared to previous innovations to a limited extent. Against this background, governance processes also need to be reconsidered. This is because the actions and effects associated with AI are no longer directly caused and controlled by human activities, but by self-learning autonomous systems that seem like a black box to humans. This leads to a high degree of uncertainty and unpredictability of AI actions and their consequences. Especially when the control mechanisms themselves become more and more technological and AI systems in an ecosystem increasingly control themselves in the future. In other words, previous development paths will be overridden and the modus operandi of governance will change.

Experiences from previous innovations and disruptions can provide inspiration and orientation for the design of an AI governance system, but are of rather limited use for the anticipation of specific AI governance requirements and measures. Given this situation, combined with potential threats to society and its ideals from AI, there are calls for a proactive governance approach to AI: that is, an approach that does not wait for challenges and then reacts with governance responses, but rather pre-emptively constrains an AI ecosystem in its development. However, caution must be exercised, as overly strict regulation will hinder the further development of AI and the exploitation of its opportunities and innovation potential (Reed, 2018; Thierer et al., 2017). Thus, policymakers and politicians face the challenge of balancing the fine line between promoting innovation and restricting AI in terms of ensuring humanity and public safety.

Closing the governance gap requires the interlocking and orchestration of governance measures and expertise from different stakeholders in all affected areas. AI governance cannot just be about AI. It combines different enabling technologies (e.g., big data, cloud computing, etc.) and related domains under one umbrella and thus affects different governance sub-areas, such as data management and protection or media regulation.

References

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management 49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review 84* (4), 1–11.

Bataller, C., & Harris, J. (2016). *Turning artificial intelligence into business value today*. (Accenture Institute for High Performance). https://www.accenture.com/nz-en/insight-artificial-intelligence-business-value.

Bennett, W. L., & Pfetsch, B. (2018). Rethinking political communication in a time of disrupted public spheres. *Journal of Communication 68* (2), 243–253. https://doi.org/10.1093/joc/jqx017.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press.

Boyd, M., & Wilson, N. (2017). Rapid developments in artificial intelligence: How might the New Zealand government respond? *Policy Quarterly 13* (4), 36–46. https://doi.org/10.26686/pq.v13i4.4619.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and Information *Technology 15* (3), 209–227. https://doi.org/10.1007/s10676-013-9321-6.

Buiten, M. C. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation 10* (1), 41–59. https://doi.org/10.1017/err.2019.8.

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration 49* (7), 751–761. https://doi.org/10.1177/0275074019856123.

Burch, S., Gupta, A., Inoue, C. Y., Kalfagianni, A., Persson, Å., Gerlak, A. K., . . . Zondervan, R. (2019). New directions in earth system governance research. *Earth System Governance 1*, 100006. https://doi.org/10.1016/j.esg.2019.100006.

Calo, R. (2010). Peeping HALs: Making sense of artificial intelligence and privacy. *European Journal of Legal Studies 2* (3), 168–192.

Calo, R. M. (2012). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 187–201). MIT Press.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine 9* (2), 48–57. https://doi.org/10.1109/MCI.2014.2307227.

Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC News*, December 2, 10.

Christensen, C. M., & Rosenbloom, R. S. (1995). Explaining the attacker's advantage: Technological paradigms, organizational dynamics, and the value network. *Research Policy 24* (2), 233–257. https://doi.org/10.1016/0048-7333(93)00764-K.

Clarysse, B., Wright, M., Bruneel, J., & Mahajan, A. (2014). Creating value in ecosystems: Crossing the chasm between knowledge and business ecosystems. *Research Policy 43* (7), 1164–1176. https://doi.org/10.1016/j.respol.2014.04.014.

Clifford, C. (2018). Google CEO: AI is more important than fire or electricity. https://www.cnbc.com/2018/02/01/googleceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html.

Coglianese, C. (2018). Optimizing regulation for an optimizing economy. *University of Pennsylvania Journal of Law and Public Affairs 4*, 1.

Coglianese, C., & Lampmann, E. (2021). Contracting for algorithmic accountability. *Administrative Law Review Accord 6*, 175.

Coglianese, C., & Lehr, D. (2016). Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ 105*, 1147.

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review 71*, 1.

Corbett, A. (2017). A systems approach to regulatory excellence. *Achieving Regulatory Excellence*, 256–270.

Dafoe, A. (2018). AI governance: A research agenda. Governance of AI Program, Future of Humanity Institute. University of Oxford.

Doneda, D., & Almeida, V. A. (2016). What is algorithm governance? *IEEE Internet Computing 20* (4), 60–63. https://doi.org/10.1109/MIC.2016.79.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., . . . Eirug, A. (2019). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. NYU School of Law, Public Law Research Paper, 20–54.

European Commission. (2019). Ethics guidelines for trustworthy AI: High-level expert group on artificial intelligence. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

European Parliament Committee on Legal Affairs. (2018). Civil law rules on robotics: European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). *Official Journal of the European Union 61*, 239–257.

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences 376* (2133), 20180081. https://doi.org/10.1098/rsta.2018.0081.

Future of Life Institute. (2017). Asilomar AI Principles. https://futureoflife.org/ai-principles/.

Future of Life Institute. (2021). National and International AI Strategies. https://futureoflife.org/national-international-ai-strategies/.

Gasser, U., & Almeida, V. (2017). A layered model for AI governance. *IEEE Internet Computing 21* (6), 58–62. https://doi.org/10.1109/MIC.2017.4180835.

Google (2021). Artificial intelligence at Google: Our principles. https://ai.google/principles/.

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt Journal of Entertainment & Technology Law 20* (2), 385–456. https://doi.org/10.31228/osf.io/5at2f.

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics 15* (2), 81–93. https://doi.org/10.1080/19331681.2018.1448735.

Iansiti, M., & Levien, R. (2004). Strategy as ecology. *Harvard Business Review 82* (3), 68–78.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly 37* (3), 101493. https://doi.org/10.1016/j.giq.2020.101493.

Jelinek, T., Wallach, W., & Kerimi, D. (2021). Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. *AI and Ethics 1*, 141–150. https://doi.org/10.1007/s43681-020-00019-y.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1* (9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons 62* (1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004.

Langer, P. F. (2020). The Formation of social credit systems—Profiling, reputation scoring, social engineering. In *Proceedings of dg.o. 2020: 21st Annual International Conference on Digital Government Research* (dg.o 2020), June 15–19, 2020, Seoul, NY, USA. ACM. http://doi.org/10.1145/3396956.3396962.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition 6* (4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008.

Linkov, I., Trump, B., Poinsatte-Jones, K., & Florin, M.-V. (2018). Governance strategies for a sustainable digital world. *Sustainability 10* (2), 440. https://doi.org/10.3390/su10020440.

Lumineau, F., Wang, W., & Schilke, O. (2021). Blockchain governance—A new way of organizing collaborations? *Organization Science 32* (2), 500–521. https://doi.org/10.1287/orsc.2020.1379.

Marchant, G. (2019). "Soft law" governance of artificial intelligence. *AI PULSE*.

Marchant, G., Abbott, K., & Allenby, B. R. (2013). *Innovative governance models for emerging technologies*. Edward Elgar Publishing. https://doi.org/10.4337/9781782545644.

Marchant, G. E. (2011). The growing gap between emerging technologies and the law. In G. E. Marchant, B. R. Allenby, & J. R. Herkert (Eds.), *The international library of ethics, law and technology. The growing gap between emerging technologies and legal-ethical oversight* (vol. 7, pp. 35–44). Springer Netherlands. https://doi.org/10.1007/978-94-007-1356-7_2.

Microsoft. (2021). Responsible AI. https://www.microsoft.com/en-us/ai/responsible-ai?.

Mitchell, A. M., Dokei, T., Hickman Tim, & Albagli, D. (2020). Japan: Regulation of artificial intelligence in Europe and Japan. https://www.mondaq.com/new-technology/1010204/regulation-of-artificial-intelligence-in-europe-and-japan.

Moore, J. F. (1993). Predators and prey: A new ecology of competition. *Harvard Business Review 71* (3), 75–86.

Moore, J. F. (1996). *The death of competition: Leadership and strategy in the age of business ecosystems*. Wiley.

Nyhan, B., & Reifler, J. (2015). The effect of fact-checking on elites: A field experiment on U.S. state legislators. *American Journal of Political Science 59* (3), 628–640. https://doi.org/10.1111/ajps.12162.

Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal 23* (8), 707–725. https://doi.org/10.1002/smj.249.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology 20* (1), 5–14. https://doi.org/10.1007/s10676-017-9430-8.

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 376* (2128), 1–12. https://doi.org/10.1098/rsta.2017.0360.

Rehm, G. (2018). An infrastructure for empowering internet users to handle fake news and other online media phenomena. In G. Rehm & T. Declerck (Eds.), *Lecture notes in computer science. language technologies for the challenges of the digital age* (vol. 10713, pp. 216–231). Springer International Publishing. https://doi.org/10.1007/978-3-319-73706-5_19.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd edition, Global edition). *Always learning*. Pearson.

Scaringella, L., & Radziwon, A. (2018). Innovation, entrepreneurial, knowledge, and business ecosystems: Old wine in new bottles? *Technological Forecasting and Social Change 136*, 59–87. https://doi.org/10.1016/j.techfore.2017.09.023.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risk, challenges, competencies, and strategies. *Harvard Journal of Law & Technology 29* (2), 354–400.

Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology 16* (3), 284–307. https://doi.org/10.2307/1930070.

Thierer, A. D., Castillo O'Sullivan, A., & Russell, R. (2017). Artificial intelligence and public policy. *Mercatus Research Paper*, 1–57. https://www.mercatus.org/system/files/thierer-artificial-intelligence-policy-mr-mercatus-v1.pdf.

Vandaele, D., Rangarajan, D., Gemmel, P., & Lievens, A. (2007). How to govern business services exchanges: Contractual and relational issues. *International Journal of Management Reviews 9* (3), 237–258. https://doi.org/10.1111/j.1468-2370.2007.00212.x.

Villasenor, J. (2020). *Soft law as a complement to AI regulation*. Brookings.

Walch, K. (2020). AI laws are coming. https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/?sh=746cdbbda2b4.

Wallach, W., & Marchant, G. E. (2018). An agile ethical/legal model for the international and national governance of ai and robotics. Paper at the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) 2018. https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_77.pdf.

Wallach, W., & Marchant, G. (2019). Toward the agile and comprehensive international governance of ai and robotics. *Proceedings of the IEEE 107* (3), 505–508. https://doi.org/10.1109/JPROC.2019.2899422.

Weyerer, J. C., & Langer, P. F. (2019). Garbage in, garbage out: The vicious cycle of ai-based discrimination in the public sector. In Y.-C. Chen, F. Salem, & A. Zuiderwijk (Eds.), *20th annual international conference on digital government research on digital government research* (pp. 509–511). ACM Press. https://doi.org/10.1145/3325112.3328220.

Weyerer, J. C., & Langer, P. F. (2020). Bias and discrimination in artificial intelligence. In I. Lee & R. Luppicini (Eds.), *Advances in e-business research. Interdisciplinary approaches to digital transformation and innovation* (pp. 256–283). IGI Global. https://doi.org/10.4018/978-1-7998-1879-3.ch011.

Wiesenberg, M., & Tench, R. (2020). Deep strategic mediatization: Organizational leaders' knowledge and usage of social bots in an era of disinformation. *International Journal of Information Management 51*, 102042. https://doi.org/10.1016/j.ijinfomgt.2019.102042.

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society a: Mathematical, Physical & Engineering Sciences 376* (2133), 1–13. https://doi.org/10.1098/rsta.2018.0085.

Wirtz, B. W., Langer, P. F., & Fenner, C. (2021). Artificial intelligence in the public sector: A research agenda. *International Journal of Public Administration 44* (13), 1103–1128.

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review 21* (7), 1076–1100. https://doi.org/10.1080/14719037.2018.1549268.

Wirtz, B. W., & Müller, W. M. (2022). An integrative collaborative ecosystem for smart cities: A framework for organizational governance. *International Journal of Public Administration*. https://doi.org/10.1080/01900692.2021.2001014.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration 42* (7), 596–615. https://doi.org/10.1080/01900692.2018.1498103.

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2022.101685.

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration 43* (9), 818–829. https://doi.org/10.1080/01900692.2020.1749851.

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons 61* (6), 823–832. https://doi.org/10.1016/j.bushor.2018.07.001.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance 2* (4), 301–313. https://doi.org/10.1093/ppmgov/gvz014.

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4* (3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.

CHAPTER 21

GOVERNING AI SYSTEMS FOR PUBLIC VALUES

Design Principles and a Process Framework

YU-CHE CHEN AND MICHAEL AHN

Introduction

Artificial intelligence (AI) governance in the public sector is a salient area of study (Zuiderwijk et al., 2021). Artificial intelligence has engendered a positive gain in efficiency and effectiveness in public service such as citizen information service, public safety, regulation enforcement, and cybersecurity (Engstrom et al., 2020; Government Accountability Office, 2018). Additionally, the use of artificial intelligence could improve government decision making for resource allocation and identify administrative errors for increasing efficiency and effectiveness (Valle-Cruz et al., 2021; Young et al. 2021b). However, applications of artificial intelligence have raised some concerns about biases against racial minorities (Angwin et al., 2016; Fountain, 2021; Harrison & Luna-Reyes, 2020) and the erosion of the administrative discretion of government employees due to automation (Barth & Arnold, 1999). A process of AI governance involving various stakeholders is expected to play an important role in ensuring that the benefits of AI can be realized while mitigating the negatives. Among these stakeholders, government has the necessary mandate to safeguard public values and to design and implement the process of AI governance.

The existing frameworks on AI governance include three approaches. The first approach consists of frameworks that connect various sectors of the society at multiple levels (societal, policy, and organization) to govern the design and application of AI systems, such as the works by Wirtz et al. (2020) and by Wirtz and Müller (2019). Here, a productive governance arrangement is assumed to require inputs from various sectors, multiple levels, and their relationships. The second approach focuses on the characteristics of specific tasks at hand and the division of authority between humans and AI, and the levels of responsibility inside a public sector organization, such as the works by Bullock (2019) and Young et al. (2019). An effective governance arrangement would require the analysis of uncertainty and complexity surrounding the administrative tasks, the relative cognitive strengths of AI in decision-making, and the requirements for the AI implementation at various levels within

the public organization. The third approach views the AI governance from a socio-technical system perspective with the recognition of the intertwined nature of social and "materials," such as the works by Janssen and Kuk (2016) and Janssen et al. (2020). In other words, an effective AI governance would require capturing the interactions between (a) societal values, relationships, and processes and (b) the technical characteristics of AI systems.

However, the existing frameworks are limited in connecting government AI systems to public values, especially from a process perspective at the level of individual systems that have a specific AI implementation for a defined public service. This chapter aims to contribute to the existing literature by (a) an explicit focus on public value creation, (b) a process orientation affording preventive solutions and phase-specific considerations, and (c) operational-level recommendations for government AI system development and implementation. A normative goal of government AI systems is to advancing public values, and these values in turn become the main criteria for assessing the impact of AI systems. A process perspective of AI governance takes a more proactive approach to creating public values and minimizing negative impacts. A process orientation also provides opportunities to understand when and how public values are introduced, codified, and implemented, as well as ensuing impacts. Such understanding creates the need to insert "levers" in the AI development and evaluation process that will have substantial impact on the performance of the AI systems in public value creation. Individual AI systems refer to AI systems that perform a specific public service or function, such as government welfare fraud detection, natural language processing for government service chatbot, and cybersecurity threat analysis and intelligence. The focus on these individual AI systems provides a more practical and operational approach to integrating public value consideration into AI and the inclusion of governance design principles aims to improve the AI system's adaptability and accountability.

This chapter contributes to our knowledge about artificial intelligence and public values by articulating human-centered, stakeholder-focused, and lifecycle-scoped as guiding governance design principles. The proposed process framework includes specific recommendations for each of the four phases of the development and implementation of AI systems: (1) goal setting; (2) iterative development decisions on data, models, and results; (3) decisions on public service; and (4) the assessment of the impacts made by the public AI systems (PAIS). Furthermore, this chapter argues for the importance of transparency and salience of diverse stakeholder participation for all phases of AI governance.

Public AI Systems and Governance Principles

Public Artificial Intelligence Systems

The primary focus of this chapter is on AI systems in the public sector. These AI systems are defined in this chapter as public artificial intelligence systems (PAIS) to entail the guiding role of public values. The pursuit of public values is a distinctive feature that separates the use of AI for profit from that of AI for public purposes. Scholars have aligned the discretionary actions augmented by artificial intelligence with public values (Barth & Arnold,

1999; Bullock, 2019; Busch & Henriksen, 2018). Public values are central to public administration and governance (Bryson et al., 2017; Nabatchi, 2018) as these values guide the purposes of government, public institutions, and organizations involved in public governance and service.

A growing recognition in the study of information technology in government is that a broader set of public values should be incorporated into consideration (Bannister & Connolly, 2014; Cordella & Bonina, 2012; MacLean & Titah, 2022). The existing literature on e-government suggests a narrow focus on managerial and service values such as efficiency, effectiveness, productivity, and client satisfaction (MacLean & Titah, 2022; Pang et al., 2014). However, such a narrow focus is problematic given the scope and variety of public values concerning our society and government. These public values include societal values such as equity, justice, fairness, equality of access, impartiality; service-oriented public values such as efficiency, effectiveness, transparency, and responsiveness; and duty-oriented values such as accountability and democratic governance (Bannister & Connolly, 2014).

Artificial intelligence has an impact on a broad set of public values classified by Bannister and Connolly (2014). These include efficiency and effectiveness, which are central to AI-enabled public service (Bullock, 2019; Shneiderman, 2020; Young et al., 2019). Furthermore, transparency of artificial intelligence is critically important for public service (Coglianese & Lehr, 2019; Larsson, 2020) and for ensuring accountability (De Bruijn et al., 2021). AI also has an impact on duty-oriented public values such as responsibility to citizens (König & Wenzelburger, 2020) and compliance with laws (Larsson, 2020). Socially oriented values impacted by government AI systems include equity (Calo, 2018; Kerr et al., 2020), accountability (Davenport & Kalakota, 2019; Young et al., 2019), privacy (Calo, 2018; Wirtz et al., 2020), and safety (Vanderelst & Winfield, 2018).

The notion of publicness in PAIS considers public value creation as the system's central goal. The creation of public values requires a governance process to identify and design the development, sources, mechanisms, and capabilities of the AI systems. For public value creation, both the involvement of key stakeholders and a structured process of governance are productive (Bryson et al., 2017). Additionally, the use of specific metrics and tools is necessary to track progress (Moore, 1995, 2013). These insights into the structure and process of public value creation enable the development of a governance framework designed specifically for AI systems with a public mission.

Governance design principles for public AI systems

Design principles are important for ensuring the effectiveness of the governance framework. For instance, Emerson and Nabatchi highlight in their work (2015) the development of joint action capacity as one of the principles for effective collaborative governance regimes. The principles offer guidance for the essential conditions and/or mechanisms for successful governance while maintaining flexibility of the specific governance arrangements. Some of the principles of AI governance appear in policy documents and governance proposals such as the OECD Council Recommendation on Artificial Intelligence (OECD, 2019). The development of principles in this chapter broadly draws from the fields of public administration, public policy, and management information system to adapt to public AI systems.

First, human-centeredness is one of the core design principles for public AI systems. Being human-centered means that an AI system will put human values and aspirations as the top priority. The European strategic document and the United States AI strategy have both articulated the need for a human-centered design. A human-centric principle guides developers and users to address the magnification of inequalities that comes with AI increasingly underpinning economic and social activities (European Political Strategy Centre, 2018). At the level of individual AI systems such a principle helps balance the growing influence of technical logic, especially in the highly technical field of AI. Technical logic favors standardization and simplicity while undermining the personalization and complexity needed for supporting autonomy and individual preferences of humans (Peeters & Widlak, 2018).

Another key governance principle is stakeholder-focused participation in the governance of public AI systems. For AI systems serving the public purposes (society as a whole), stakeholders include individuals and organizations in government, industry, nonprofit, academia, and civil society. This set of stakeholders is broader than that of AI systems created and serviced by technology corporations with a focus on satisfying customer needs and the creation of values for the corporation and its shareholders. Therefore, citizens should serve as partners in co-producing personalized and general public services by interacting with algorithm-powered systems (Williamson, 2014). Governments, representing the interest of the public, need to play a key role in defining public interest, providing funding for public missions, and establishing governance rules and regulations. Academia, as a relatively independent research operation, is a key stakeholder tasked with focusing on public missions and advancing public values. For instance, the Stanford Institute for Human-centered Artificial Intelligence has served as an academic focal point for integrating the expertise from academia, industry, and government via research, workshops, conferences, and policy reports (e.g., Stone et al., 2016). The Institute launched the impact of AI as one of the three core areas of research. It also led the creation and annual update of AI index to systematically measure the implications of AI for economy, education, ethics, diversity, and national strategies.

Moreover, populations directly impacted by the public service decisions made by AI should be represented or consulted during AI's developmental and implementation phases. The technical complexity and knowledge requirements for understanding AI have created a widening divide in understanding the development and implementation of AI for public service. This widening digital divide tends to weaken the position of already disadvantaged populations that are directly being impacted by the service decisions made by an AI-enabled algorithm.

Another governance design principle is to adopt a lifecycle approach to the governance of public AI systems to understand how public values are introduced, codified, implemented, and impacted. A lifecycle approach considers the entire lifetime of a product, process, or technology from the design and production to distribution, use, and disposal (Guinée et al., 2011; Tukker, 2000). Additionally, a lifecycle approach focuses on purposes and substantive outcomes (Cashmore, 2004). For public AI systems, an application of the lifecycle principle would require an emphasis on energy and data inputs, societal impact, and ethical concerns pertinent to the entire lifecycle of development, use, and management. For computationally intensive AI systems, energy input is an important factor, along with data storage, access, and manipulation. Societal impacts and ethical concerns that have been

raised about AI should be an integral part of impact assessment. These social impacts and ethical concerns include, but are not limited to, the autonomy of administrative staff and equality in treating service recipients. This lifecycle impact assessment approach, when adapted to PAIS, needs to capture the distinctive focus on data, process, methodology, and code for a particular application. Moreover, the lifecycle approach involves the need to assess the impact of unauthorized access to and the malicious use of an algorithm as well as the unintended consequences of algorithmic decision-making based on codes developed by machine learning.

The application of the lifecycle principle for AI governance can also bring some more perpetual and exponential impacts of technology to the forefront of societal attention. Some of the codes and processes for machine learning could continue to exist on a code development and repository platform (e.g., GitHub). The access and use management of these algorithms developed by AI systems for public purposes are critical for understanding long-term AI ecosystem-level impacts. Making these algorithms and documentation available is listed as one of the national strategies for AI development in the United States (Select Committee on Artificial Intelligence of the National Science & Technology Council, 2019). Furthermore, the lifecycle impact assessment approach allows integration with an emphasis on stakeholder participation as seen in the U.S. National Environmental Protection Act. A similar structure and process can also work for PAIS to integrate stakeholders impacted by the development and use of PAIS.

A Process Governance Framework for Public AI Systems

This process framework aims at the level of individual AI systems for analytical purposes while incorporating broader technical and social processes at the organizational and societal levels. As depicted in Figure 21.1, the suggested four phases are: (1) goal setting; (2) iterative development decisions on data, models, and results; (3) decisions on public service; and (4) impacts of these public service decisions. This process framework situates AI systems in the public sector with explicit consideration of the public values as goals and purposes of AI systems as well as the impacts of AI-enabled public service decisions, in addition to AI system development and use. Each phase of the process framework identifies the associated governance challenge for public value creation. As summarized in Table 21.1, the discussion of each phase identifies the emphasis of salient design principles, the mechanism of public value creation, and aspects of transparency.

Goal setting: Introduction of public values

The first phase for a public AI system is goal setting, in which the main public values are introduced to guide the rest of development and implementation. Setting the goals of AI-augmented digital assistance services can introduce public values such as efficiency and effectiveness (Androutsopoulou et al., 2019; Aoki, 2020; Borfitz, 2019). To achieve these goals,

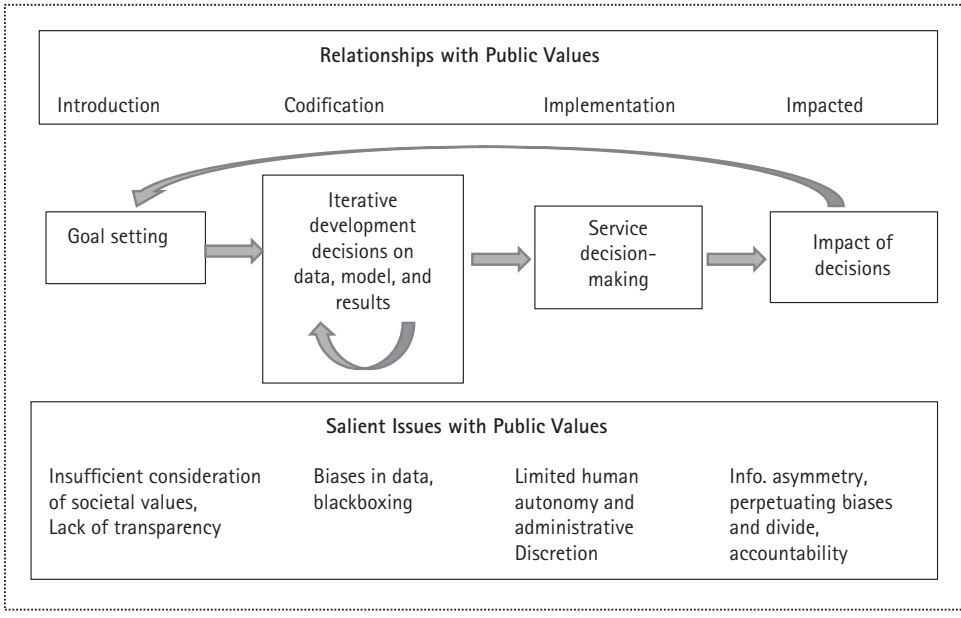

FIGURE 21.1 Process Perspective of the Relationships between AI Systems and Public Values

an AI-empowered cognitive assistance system could support human decision-makers in information processing and preference ranking in decision-making. Equity can be introduced as one of the goals in the form of serving citizens with varying capabilities. For instance, AI can provide visually impaired and elderly people with viable alternatives (e.g., voice-based navigation) to interact with government systems and multi-lingual options for people speaking different native languages (Eggers et al., 2017, p. 12).

One governance challenge at this phase is the tendency to narrowly focus on efficiency and effectiveness of public service provision. A broader consideration of public values that are societal-oriented or duty-oriented as articulated by Bannister and Connelly (2014) is lacking. This focus on service efficiency and effectiveness is evident in the references documenting the applications of AI (e.g., Desouza, 2018; Eggers et al., 2017; Government Accountability Office, 2018). Examples include efficiency and effectiveness in prioritizing public safety resources and in addressing cybersecurity (Government Accountability Office, 2018; Wirtz et al., 2019). Another governance challenge at this phase is the lack of transparency to people outside a government agency regarding public value identification and assessment for a particular AI system. Lack of transparency is a notable governance challenge, particularly for AI (Adadi & Berrada, 2018; Reddy et al., 2020). Transparency on the specific public values chosen as goals of the AI system is crucial to understanding how public values interact with AI systems.

A governance body needs to be in place for the explicit consideration of a wide range of public values. Considering broad societal values is important for the governance of AI (Larsson, 2020; Reddy et al., 2020; Wirtz et al., 2020). Society-oriented public values such as fairness and justice can be incorporated as ethical principles to prevent bias or discrimination of AI systems (Larsson, 2020; Winfield et al., 2019). For public sector organizations, a process requirement for considering a wide range of public values as goals

Table 21.1 Phase-specific Governance Principles, Public Value Creation, and Transparency

	Goal setting	Iterative development decisions on data, model, and results	PAIS service decision-making	Impacts of PAIS decisions
Emphasis of design principles	Stakeholder-focused involvement, lifecycle approach	Human-centered social logic and understanding	Human-centered artificial discretion arrangement, stakeholder-focused staff and client involvement	Human-centered evaluation of stakeholder-focused impacts, lifecycle approach
Creation of public values	Articulation and inclusion of societal values with scorecards	Adaptation of social logic as the development logic, critical review of biases	Promotion of human autonomy to generate trustworthy human-machine discretion, identification, and correction of bias	Advancing efficiency and effectiveness, attention to societal values (equity, accountability, privacy, etc.)
Transparency (on what and how)	Transparency on public values as AI system goals, transparency on stakeholder participation, metrics of value impacts	Explainable AI focusing on known and potential bias	Transparency on scope, authority, and reasoning of AI-based decision-making, transparency on the participation of impacted populations	Transparency on the stakeholder-specific public value impacts, distributional effect among stakeholders via evaluation
Transparency (to whom)	Policymakers, representatives of the impacted populations, technical experts	Government administrative staff, government technical staff, service recipients, and independent technical experts	Government decision-makers, impacted population groups pertinent to a type of public service decisions	Impacted populations, stakeholders in the public, private, and nonprofit sectors

in addition to efficiency and effectiveness could be helpful. A governance body can take many different forms such as an advisory board or an oversight committee (Winter & Davidson, 2019).

The principle of stakeholder involvement is key at this phase. For AI systems serving the public, stakeholders include individuals and organizations in government, industry, nonprofit, academia, and civil society (Mikhaylov et al., 2018). Citizens, as one key stakeholder group, should play the role of partners with government in co-producing public service by interacting with algorithm-powered systems (König & Wenzelburger, 2020; Williamson, 2014, 2015). The stakeholders tend to have a different set of values as connected to a particular

AI system (Dwivedi et al., 2021; Sun & Medaglia, 2019). More specifically, stakeholder involvement should be broad-based to include policymakers and administrative staff using the system along with AI developers. Representatives of the populations impacted by the specific AI use need to be involved in the goal-setting phase because the participation of the impacted populations ensures consideration of a broader set of public values that includes equality and autonomy. Technical experts constitute an important stakeholder group that can incorporate technical knowledge to help understand the process of AI development and its potential malicious use.

The application of a lifecycle principle is important at the phase of goal-setting. The consideration of the complete lifecycle of AI system development and implementation offers opportunities to develop preventive measures against bias being codified and implemented to cause harm to the impacted populations. For instance, equity can serve as an explicit goal that guides the development of AI systems to prevent adverse equity impact at the public service decision phase. In addition, a lifecycle approach incorporates the learning from the evaluation of public value impacts on similar AI systems. The results of these evaluations can inform the key stakeholders of the unintentional consequences of AI systems on public values. Such information helps best incorporate the consideration of these consequences at the first phase of goal setting for public value creation.

At this phase, transparency is crucial. In aligning involvement from broad-based stakeholders, transparency will allow them to articulate the specific public values that need to be introduced and the rationale for advancing these values. Transparency about the entire process of stakeholder participation, including who participates and what their opinions are, is essential. Equally important regarding transparency is how the stakeholders measure and track the advancement of public values to ensure accountability.

Iterative development decisions on data and models: Codification of public values

The second phase includes the iterative development decisions on data and models for AI-enabled algorithms. The decision on data use for AI-enabled algorithm development introduces and codifies public values. For public service, data carry institutional history, policy focus, and potential bias towards certain segments of the population (Janssen et al., 2020; Janssen & Kuk, 2016). Without an explicit review, biases in data against the realization of public values are embedded and formalized in algorithm development. An AI system could focus on efficiency as a public value for policing resource use (Ku & Leroy, 2014). However, efficiency in policing can become profiling and targeting minorities when there are biases in the data against these minorities (Angwin et al., 2016).

A public value audit of both data and AI-enabled algorithms needs to first check their alignment with the intended public values established at the first goal-setting phase. This audit needs to pay attention to equality, fairness, and accountability, as well as to other known biases. Data auditing is important to check if there are inherent biases in the data (Henman, 2020; Janssen et al., 2020; Winter & Davidson, 2019). Public service data are rooted in a particular historical and socioeconomic context with specific public policy objectives. As a result, failing to critically review the implicit bias in the data, model, and

results could have the consequence of perpetuating an existing bias (Angwin et al., 2016; Boyd & Crawford, 2012).

Similar audits should also be conducted on potential biases in the algorithm developed from the data and modeling choices. The bias in the modeling choice is subtler, but it mostly resides in the appropriateness of the models selected to approximate the phenomenon in question. If the understanding of the phenomenon has a built-in ethical bias, the model chosen as the basis of learning could further amplify that bias. Such auditing, if aided by a human-centered logic, could help with the incorporation of societal concerns and ethical standards (Winfield & Jirotka, 2018).

The choice of models and the goal of optimization could also affect how values are introduced and reinforced. The choice of model tends to amplify a data-driven technical logic and these technical choices can reinforce the technical logic of standardization and simplification at the expense of thorough value discussion and balance (Cordella & Tempini, 2015). Such technical logic for the optimization of an algorithm tends to narrowly operationalize the public value in question. For instance, the technical focus of algorithm development would focus on the predictive power of algorithms to identify potential high crime risk transportation areas without fully considering the ethical implications of such identification (Kouziokasa, 2017). With such a technical focus, much less would be discussed about public value implications for such predictions.

Following the human-centered governance principle can help balance the over-reliance on technical logic that tends to focus on efficiency and effectiveness of the technology. Technical logic such as functional simplification tends to play an important role in the public-sector use of information technology (Cordella & Tempini, 2015). The technically demanding nature of AI is likely to reinforce the focus on technical logic and, consequently, on efficiency and effectiveness. Therefore, a conscious effort needs to be undertaken to incorporate social logic such as the consideration of societal values as embodied in ethics (Winfield et al., 2019) and unique circumstances surrounding a decision to serve as a balancing force.

The technical nature and machine learning aspect of this AI development phase have created a transparency challenge for human understanding (Adadi & Berrada, 2018; Arrieta et al., 2020; König & Wenzelburger, 2020). Transparency is critical at this phase with the goal of engendering human trust in the prediction, classification, and recommendations by algorithms via machine learning. Machine learning has continued to grow in the complexity of data and the sophistication of computational techniques (Arrieta et al., 2020). Moreover, the technical nature of AI development creates a black box that is hard to decipher because of the additional consideration and effort involved in the documentation and translation of machine-learning into its why and how in language that can be understood by humans.

A governance recommendation for transparency includes the stakeholders who should be informed, as well as the nature and communication of the content. At this phase of development, four groups are central to effective public value audits. The first group constitutes the administrative staff inside the public organization who have knowledge about the public value goals of an AI system and administrative data. The second group consists of the technical staff who are making choices about data and models. The third group pertains to the service recipients. The participation of these recipients could be accomplished through active consultation. The final group consists of independent technical experts who are aware of the standards and practices of safeguards against biases in data and modeling.

Transparency also entails making AI explainable to various stakeholders to make AI systems responsible and accountable (Adadi & Berrada, 2018; Arrieta et al., 2020; König & Wenzelburger, 2020). Explainable AI at this phase can focus on clarifying how machine learning transpires and why the developed algorithm arrives at a particular result. Explanations on the type of machine learning used and its level of sophistication are critical. The goal is to achieve explainable AI calibrated for a diversity of knowledge levels. The advances in explainable AI have offered some insights and recommended practices (Arrieta et al., 2020). In terms of content, the technical nature of the data selection, model choices, and algorithm-based results would require the establishment of templates and guidelines for human understanding, especially with various levels of technical knowledge and training. The explainability of the technical knowledge would be essential for effective communication. For PAIS, an explanation of purpose, function, and rationale for development decisions is also crucial (Coglianese & Lehr, 2017; Turek, 2016). Such communication needs to offer explanations of the rationales for learning, reasoning, and generation of results suitable for human consumption.

Public service decision-making

The third phase involves decision-making about services that are enabled by public AI systems. This phase addresses how public values are being implemented through the production and delivery of public service with the collaboration between humans and AI systems. In the public sector, such collaboration is structured by respective authority rules on who has the authority to make what decisions and the level of automation by AI in making those decisions. The higher the level of automation in decision-making based on AI-generated algorithms, the more influence AI systems have over a decision. Increasing scope of automating public service decisions is evident for system-level bureaucrats (Barth & Arnold, 1999; Bullock, 2019; Young et al., 2019) and for national government agencies (Engstrom et al., 2020). The growing use of chatbots for offering public services allows AI systems to make decisions on the relevant information to provide to individuals requesting services (Androutsopoulou et al., 2019).

Human autonomy is one of the main values of concern at this phase because the level of automation undertaken by AI systems impacts human autonomy in making decisions. As the imbalance of cognitive capabilities between humans and AI systems continues to widen, human autonomy in decision-making is likely to diminish. More specifically in the public sector, government employees should have the autonomy to exercise their discretion given their authority. However, the introduction of AI could accelerate the erosion of government employees' discretionary authority (Bullock et al., 2020). Even when government employees have the final authority, it is cognitively infeasible for them to challenge the recommendations made by AI systems without a high level of explainability of the decisions. AI systems could further undermine the autonomy of humans when such systems create humans' complacency and dependence on them (Parasuraman & Manzey, 2010).

An effective governance mechanism at this phase needs to support a human-centered governance principle for utilization of AI systems. A human-centered approach discerns the conditions under which human discretion is more suitable and then supports human

autonomy accordingly. The study of artificial discretion in the context of public administration has explored uncertainty and complexity as main criteria for deciding the division of discretionary authority (Bullock, 2019; Young et al., 2019). AI should play the main role in decision-making when the task has low uncertainty (high analyzability) and low complexity (few deviations). Humans should be the lead decision-makers when the task is high on both complexity and uncertainty (Bullock, 2019). A human-centered governance principle implies a preference for giving discretionary authority to administrative staff.

At this phase, society-oriented public values, such as equality in treatment and access, fairness, accountability, and due process, can be impacted by automated AI systems. Public sector organizations are responsible for advancing these society-oriented public values. These values are typically being codified in laws and rules on administrative agencies. For instance, administrative procedure laws and rules require due process and equality of treatment in the United States. The use of AI systems in government could promote fairness by reducing human bias in favoring people with demographics similar to those of administrative staff. At the same time, such use could reinforce biases codified at the previous phase of AI system development.

An application of the stakeholder involvement principle would require the participation of administrative staff and service recipients because both are directly impacted by governance design at this service decision phase. The autonomy of administrative staff in adjusting decisions based on unique circumstances is at stake when AI systems are responsible for public service production and delivery. A review of literature has suggested an increase in using technology and algorithms for street-level administrative decision-making (Busch & Henriksen, 2018). Consequently, an important step is to meaningfully engage administrative staff in governance design that aligns with the goal of AI systems along with human autonomy. Governance design simultaneously needs to engage service recipients (or their representatives) to understand the implementation of public values such as equality, accountability, and due process. The concern about biases and equity (Janssen & Kuk, 2016; Young et al., 2019; Young et al., 2021a) raises a critical need to bring in the perspective of impacted populations to assess which (and to what extent) society-oriented public values are advanced or eroded.

Transparency on service decision-making must strive to inform human decisionmakers of the scope, authority, and reasoning of AI-based decision-making. Such transparency is essential not only for the accountability of AI systems internal to decisionmakers inside a government agency but also for accountability to society (Young et al., 2021a). An explanation of rationales with the availability of supporting evidence is critical for transparency in algorithm-based administrative decision-making (Coglianese & Lehr, 2017). Human decisionmakers need to understand data, machine learning processes, and models of the AI systems to understand the why and how of the decisions rendered on a specific task. More importantly, human decisionmakers need to fully understand when an AI-enabled algorithm fails or succeeds in completing the tasks. All these explanations need to be given in a way that is understandable by humans via an explainable interface (Adadi & Berrada, 2018; Arrieta et al., 2020; Government Accountability Office, 2018, p. 19).

In addition, the role of impacted populations in decision-making needs to be transparent. The primary public value of concern is equity, regarding the level of transparency

about the extent to which the stakeholders' concerns are incorporated into the decision-making process. For digital assistance service, inclusion could be accomplished via the collection of individual data for personalized service. For decisions on the provision of government benefits, consultation with front-line service providers and service recipients could be conducted. The public commenting process for regulations can serve as a possible mechanism for making the respective authority of decisionmakers (humans vs AI) transparent to stakeholder groups impacted while also seeking their input.

Impacts of PAIS on public values

The final phase pertains to the impact of decisions via a PAIS made primarily on public values. Development of methodologies for impact assessment is essential for improving our knowledge about artificial intelligence in government (Medaglia et al., 2021). At this phase, the main relationship between AI systems and public values is that of impact. The aim of this phase is to evaluate a wide range of public values. These public values include the service-oriented ones that are typically explicitly mentioned in the goals and objectives of the government AI systems, such as efficiency, effectiveness, and responsiveness. Furthermore, the public has certain concerns about the lack of responsiveness (König & Wenzelburger, 2020) and violation of laws (Larsson, 2020). Other socially-oriented values impacted by government AI systems include equity (e.g., Calo, 2018; Kerr et al., 2020; Montes & Goertzel, 2019), accountability (Butcher & Beridze, 2019; Davenport & Kalakota, 2019; Young et al., 2019), privacy (Calo, 2018; Reddy et al., 2020; Winter & Davidson, 2019; Wirtz et al., 2020), and safety (Vanderelst & Winfield, 2018). Additionally, the possible malicious uses of AI (Butcher & Beridze, 2019; Kerr et al., 2020) may not be foreseen at the earlier phases of the process.

One primary governance challenge is to address the informational asymmetry between government and citizens regarding the AI system in question. Such asymmetry is pronounced when certain errors in the data are exacerbated via algorithms developed by AI. Such errors in data later populated in various government information systems are difficult to identify in the first place and more challenging to correct and compensate afterwards (Peeters & Widlak, 2018). This difficulty is due to the level of technical expertise and the quantity of resources required for citizens to even conduct an investigation. Such inequality is likely to worsen with the use of AI that further increases the gap of technical knowledge between AI system architects and ordinary citizens. Additionally, the degree of inequality further increases when AI-enabled public service decisions are made for disadvantaged populations. These populations are likely to have limited or no access to social capital and technical expertise to counter algorithm-based decisions. Another governance challenge is associated with the cost and control that government has on the AI systems for accountability. One of the challenges that governments encounter is the high resource demand to thoroughly audit AI systems purchased from the private sector and the resulting imbalance of expertise possessed by the AI company versus the government agency (Desouza, 2018; Engstrom et al., 2020).

An integral part of the governance solution at this phase is a requirement to conduct public value impact analysis for various stakeholder groups, with the focus on disadvantaged populations being impacted by the AI system. Such an approach follows the

human-centered and stakeholder-focused governance design principles. This approach is particularly relevant given the negative impact of algorithm-based decision-making on equity. A stakeholder-focused evaluation (Posavac & Carey, 2003) is needed to check against automating and/or amplifying inequality impacted by public AI systems. This evaluation should have a rigorous assessment framework that has been developed with representation from the impacted populations. In addition, results of the impact assessment should be reported with an explicit section on equity while addressing the impact on other public values identified at the goal-setting phase. Such reporting follows the lifecycle principle to inform the goal-setting of the same AI systems or that of other AI systems in the future.

The transparency of the distributional effect of AI systems on various stakeholders at this phase helps address the informational asymmetry between government and disadvantaged populations. Using a technically sophisticated algorithm as a means of making decisions has raised the issues of creating technical barriers for the impacted population to seek remediation. The complexity and requirements of government rules and regulations initially designed to increase accountability have created a disproportional cost and an administrative burden for disadvantaged populations to receive benefits. Therefore, the focus should be on the physically, socially, and economically disadvantaged populations that are impacted by these AI systems. Transparency needs to be communicated in a manner that enhances the understanding of the impacted disadvantaged populations.

Another component of the governance solution at this phase involves cross-sector collaboration that addresses information asymmetry and benefits from perspectives and expertise from various sectors. Collaborative governance has been recommended as a means of addressing the broad concerns about the impacts of AI systems on our society (Lauterbach, 2019; Mikhaylov et al., 2018; Todolí-Signes, 2019). These impacts involve socially oriented public values. The scope of such collaboration goes beyond a single sector (public, private, or nonprofit) (Mikhaylov et al., 2018). The goal is to create a central knowledge repository of government AI's impacts on public values and respective governance solutions; this repository could take the form of a coordinating committee (Wallach & Marchant, 2019), which is supported by facilitative leadership for sharing knowledge, insights, and strategies (Mikhaylov et al., 2018).

An integrated governance solution

In sum, an integrated governance solution needs to address the important governance challenges for each phase, as well as connect the four phases, as illustrated in Figure 21.2. One integrated governance arrangement is to have a governance structure encompassing representatives of stakeholder groups including government administrative staff, government AI system developers, impacted citizens, independent AI experts, and citizens at large. Such an arrangement is inclusive of the main stakeholder groups across all the phases. Moreover, such a governance structure should have the mandate of considering a broad range of public values throughout all phases while being attentive to public values that are salient but less articulated. The adoption of an instrument for tracking public values throughout the process would be productive. The public value scorecard approach advanced

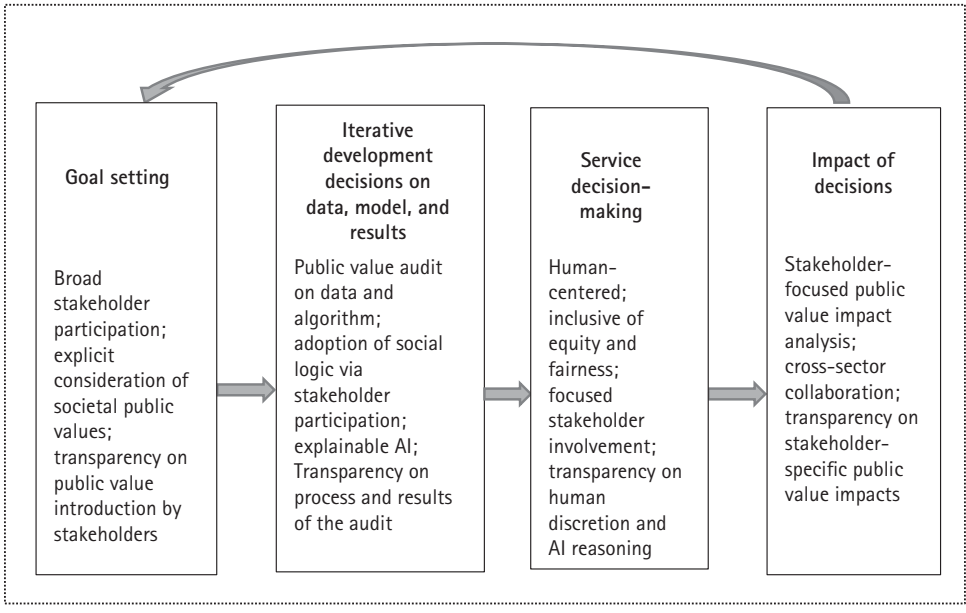
Goal setting
Broad stakeholder participation; explicit consideration of societal public values; transparency on public value introduction by stakeholders
Iterative development decisions on data, model, and results
Public value audit on data and algorithm; adoption of social logic via stakeholder participation; explainable AI; Transparency on process and results of the audit
Service decision-making
Human-centered; inclusive of equity and fairness; focused stakeholder involvement; transparency on human discretion and AI reasoning
Impact of decisions
Stakeholder-focused public value impact analysis; cross-sector collaboration; transparency on stakeholder-specific public value impacts

FIGURE 21.2 A Process-oriented Integrated AI Governance Solution

by Moore (1995) can serve as a productive mechanism for measuring the progress made on the creation of public values. Such a scorecard, as agreed upon by key stakeholders, will provide an important assessment framework throughout the process.

This governance structure needs to adopt a process perspective that considers the introduction, codification, implementation, and impact of public values. The consideration of societal public values is particularly helpful for the goal-setting phase. The codification of public values at the development phase needs to ensure the alignment of public values set at the previous phase. The implementation requires a public value audit to ensure that service decisions reflect the core public values. The final phase entails stakeholder-specific public value impact evaluation to promote both alignment and learning, which serves as the basis for feedback to public values as the goal of PAIS.

The entire process should ensure transparency to various stakeholders at all phases. Transparency on both contents and processes is essential for AI accountability (Busuioc, 2021). Transparency would require a structured reporting and communication protocol for the groups of stakeholders identified. Reporting and communication should follow a formalized plan that identifies when, with whom, and what information is shared. Transparency on public values introduced is essential for the first phase of goal setting. The second phase needs to ensure transparency on data and potential bias embedded in the data (Henman, 2020). For AI-enabled results, explanations should be given on the strengths and weaknesses of the algorithms enabled by machine learning (Turek, 2016). For the service decision-making phase, transparency on human discretion and AI reasoning would aid in understanding the implications of public value implementation in these public AI systems. For the final phase, the transparency of public value impacts is needed to be specific to a particular stakeholder group.

Conclusion

This chapter aims to advance our knowledge about the interconnections between public values and artificial intelligence at the level of AI system development and implementation. This chapter introduces public AI systems (PAIS) as the conceptual foundation for public values as the focus of value creation to address societal concerns that are central to public policy and administration. This concept of PAIS further allows for discovering governance principles and mechanisms for advancing public values at a specific phase of AI system development and implementation.

This chapter further outlines three governance principles that are designed to guide the development of a governance framework for PAIS. The human-centered principle aligns PAIS with human societal values and human autonomy in policy and service decision-making. This principle helps balance the prevalence of technical logic and machine-dominant decision-making. The stakeholder-focused principle underscores the need for stakeholder participation in PAIS development and this principle assists in discovering the bias against impacted vulnerable populations. The lifecycle principle considers the impacts on public value at each phase of PAIS development and assesses the public value implications of decisions made at earlier phases. The application of these principles at specific phases of PAIS development and implementation further instructs when and how public values are introduced, codified, implemented, and impacted.

The process governance framework consists of four specific phases: (1) goal setting; (2) iterative development decisions on data, model, and results; (3) PAIS service decision-making; and (4) impacts of these service decisions. The framework then provides a phase-specific application of salient governance principles, as well as recommendations for advancing public values and improving transparency. For goal setting, effective PAIS governance framework needs to consider a broad set of societal values and stakeholder participation. For development, governance should entail an audit on data and algorithm and on the balance between technical logic and societal concerns to ensure AI's contribution to public value.

Regarding AI system implementation, an effective governance framework follows the human-centered principle for an optimal allocation of authority between administrative staff and machine. Here, equity and fairness need to be explicitly considered and the impact of the decisions made by a public AI system should be subjected to stakeholder-specific public value impact analysis.

An integrated governance solution emerging from the process governance framework is the creation of an interdisciplinary cross-sector governance structure. This structure can be charged with overseeing the entire process to implement measures for proactively creating public values throughout the process and aligning actions with the set goals of public value creation. Furthermore, the transparency of the interaction between a specific phase of AI system development and implementation is crucial for governing AI systems. The aim of such transparency is to increase the various stakeholders' understanding of the public value implications of AI systems. Such understanding promotes meaningful stakeholder participation and mitigates the potential negative impact on equity and disadvantaged populations.

The main contribution of this chapter is a process governance framework that focuses on public values for AI systems in the public sector. The theoretical implication of the framework is the advancement of our understanding of the interaction between the phases of AI development and implementation and their relationships with public values. The process governance framework aids in integrating various bodies of literature to develop a theory of public value creation for PAIS. The practical implication is the articulation of phase-specific solutions, as well as an integrated governance framework, for the realization of public values for AI systems. Future studies can further advance our understanding of the application of the proposed process framework in addressing AI governance challenges.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access 6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052.

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly 36*(2), 358–367.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *Propublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly 37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabikg, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion 58*, 82–115.

Bannister, F., & Connolly, R. (2014). ICT, public values and transformative government: A framework and programme for research. *Government Information Quarterly 31*(1), 119–128.

Barth, T. J., & Arnold, E. (1999). Artificial intelligence and administrative discretion. *American Review of Public Administration 29*(4), 332–351.

Borfitz, D. (2019). Digital Assistants Transforming Public Service. *AItrends Blog*. July 15, 2019. https://www.aitrends.com/ai-world-government/digital-assistants-transforming-public-service/.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information Communication & Society 15*(5), 662–679.

Bryson, J., Sancino, A., Benington, J., & Sørensen, E. (2017). Towards a multi-actor theory of public value cocreation. *Public Management Review 19*(5), 640–654. https://doi.org/10.1080/14719037.2016.1192164.

Bullock, J. (2019). Artificial intelligence, discretion, and bureaucracy. *American Review of Public Administration 49*(7), 751–761.

Bullock, J., Young, M. M., & Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity: The International Journal of Government & Democracy in the Information Age 25*(4), 491–506.

Busch, P. A., & Henriksen, H. Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity 23*, 3–28.

Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293.

Butcher, J., & Beridze, I. (2019). What is the state of artificial intelligence governance globally? *The RUSI Journal 164*(5–6), 88–96. https://doi.org/10.1080/03071847.2019.1694260.

Calo, R. (2018). Artificial intelligence policy: A primer and roadmap. *University of Bologna Law Review 3*(2), 180–218. https://doi.org/10.6092/issn.2531-6133/8670.

Cashmore, M. (2004). The role of science in environmental impact assessment: process and procedure versus purpose in the development of theory. *Environmental Impact Assessment Review 24*(4), 403–426.

Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal 105*, 1147.

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review 71*(1), 1–56.

Cordella, A., & Bonina, C. M. (2012). A public value perspective for ict enabled public sector reforms: A theoretical reflection. *Government Information Quarterly 29*(4), 512–520.

Cordella, A., & Tempini, N. (2015). E-government and organizational change: Reappraising the role of ICT and bureaucracy in public service delivery. *Government Information Quarterly 32*(3), 279–286.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal 6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

De Bruijn, H., Warnier, M., & Janssen, M. (2021). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly 39*(2). https://doi.org/10.1016/j.giq.2021.101666.

Desouza, K. C. (2018). *Delivering artificial intelligence in government: Challenges and opportunities*. IBM Center for the Business of Government.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., . . . Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management 57*. https://doi.org/10.1016/j.ijinfomgt.2019.08.002.

Eggers, W. D., Schatsky, D., & Viechnicki, P. (2017). *AI-augmented government: Using cognitive technologies to redesign public sector work*. Deloitte Center for Government Insights.

Emerson, K., & Nabatchi, T. (2015). *Collaborative governance regimes*. Georgetown University Press.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by algorithm: Artificial intelligence in federal administrative agencies*. Administrative Conference of the United States.

European Political Strategy Centre. (2018). *The age of artificial intelligence*. European Commission.

Fountain, J. E. (2021). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly 39*(2). https://doi.org/10.1016/j.giq.2021.101645.

Government Accountability Office. (2018). *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications*. (GAO-18-142SP). Government Accountability Office.

Guinée, J. B., Heijungs, R., Huppes, G., Zamagni, A., Masoni, P., Buonamici, R., . . . Rydberg, T. (2011). Life cycle assessment: Past, present, and future. *Environmental Science & Technology 45*(1), 90–96. https://doi.org/10.1021/es101316v.

Harrison, T. M., & Luna-Reyes, L. F. (2020). Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review 40*(2), 494–511. https://doi.org/10.1177/0894439320980122.

Henman, P. (2020). Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration 42*(4), 209–221. https://doi.org//10.1080/23276665.2020.1816188.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly 37*(3). https://doi.org/10.1016/j.giq.2020.101493.

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly 33*, 371–377.

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society 7*(1), 205395172091593. https://doi.org/10.1177/2053951720915939.

König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly 37*(3), 101489. https://doi.org/10.1016/j.giq.2020.101489.

Kouziokasa, G. N. (2017). The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment. *Transportation Research Procedia 24*, 463–473.

Ku, C.-H., & Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly 31*(4), 534–544.

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society 7*(3), 1–15. https://doi.org/10.1017/als.2020.19.

Lauterbach, A. (2019). Artificial intelligence and policy: Quo vadis? *Digital Policy, Regulation and Governance 21*(3), 238–263. http://dx.doi.org/10.1108/DPRG-09-2018-0054.

MacLean, D., & Titah, R. (2022). A systematic literature review of empirical research on the impacts of e-government: A public value perspective. *Public Administration Review 82*(1), 23–38. https://doi.org/10.1111/puar.13413.

Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2021). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review* 089443932110340. https://doi.org/10.1177/08944393211034087.

Mikhaylov, S. J., Esteve, M., & Campion, A. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376*(2128), 20170357.

Montes, G. A., & Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change 141*, 354–358. https://doi.org/10.1016/j.techfore.2018.11.010.

Moore, M. H. (1995). *Creating public value: Strategic management in government*. Harvard University Press.

Moore, M. H. (2013). *Recognizing public value*. Harvard University Press.

Nabatchi, T. (2018). Public values frames in administration and governance. *Perspectives on Public Management and Governance 1*(1), 59–72.

OECD. (2019). *Recommendation of the council on artificial intelligence*. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Pang, M.-S., Lee, G., & Delone, W. H. (2014). IT resources, organizational capabilities, and value creation in public-sector organizations: A public-value management perspective. *Journal of Information Technology 29*(3), 187–205.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors 52*(3), 381–410. https://doi.org/10.1177/0018720810376055.

Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture—The case of the Dutch civil registry's master data management system. *Government Information Quarterly 35*, 175–183.

Posavac, E. J., & Carey, R. G. (2003). *Program evaluation: Methods and case studies*. 6th edn. Prentice Hall.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association 27*(3), 491–497. https://doi.org/10.1093/jamia/ocz192.

Select Committee on Artificial Intelligence of the National Science & Technology Council. (2019). *The national artificial intelligence research and development strategic plan: 2019 update*. Executive Office of the President.

Shneiderman, B. (2020). Bridging the gap between ethics and practice. *ACM Transactions on Interactive Intelligent Systems 10*(4), 1–31. https://doi.org/10.1145/3419764.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., . . . Teller, A. (2016, September). *Artificial intelligence and life in 2030*. Stanford University.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly 36*(2), 362–383.

Todolí-Signes, A. (2019). Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: The necessary collective governance of data protection. *Transfer: European Review of Labour and Research 25*(4), 465–481. https://doi.org/10.1177/1024258919876416.

Tukker, A. (2000). Life cycle assessment as a tool in environmental impact assessment. *Environmental Impact Assessment Review 20*(4), 435–456.

Turek, M. (2016). *Explainable Artificial Intelligence (XAI) program*. Department of Defense. https://www.darpa.mil/program/explainable-artificial-intelligence.

Valle-Cruz, D., Fernandez-Cortez, V., & Gil-Garcia, J. R. (2021). From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. *Government Information Quarterly* 39(2). http://doi.org/10.1016/j.giq.2021.101644.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research 48*, 56–66. https://doi.org/10.1016/j.cogsys.2017.04.002.

Wallach, W., & Marchant, G. (2019). Toward the agile and comprehensive international governance of AI and robotics [point of view]. *Proceedings of the IEEE 107*(3), 505–508. https://doi.org/10.1109/jproc.2019.2899422.

Williamson, B. (2014). Knowing public services: Cross-sector intermediaries and algorithmic governance in public sector reform. *Public Policy and Administration 4*(29), 292–312.

Williamson, B. (2015). Governing software: Networks, databases and algorithmic power in the digital governance of public education. *Learning, Media and Technology 40*(1), 83–105. https://doi.org/10.1080/17439884.2014.924527.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376*(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. *Proceedings of the IEEE 107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622.

Winter, J. S., & Davidson, E. (2019). Governance of artificial intelligence and personal health information. Digital Policy, Regulation and Governance 21(3), 280–290.

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review 21*(7), 1076–1100.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration 42*(7), 596–615.

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration 43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851.

Young, M., Himmelreich, J., Honcharov, D., & Soundarajan, S. (2021b). *The right tool for the job? Assessing the use of artificial intelligence for identifying administrative errors*. Paper presented at the DGO 2021: The 22nd Annual International Conference on Digital Government Research, Omaha, NE, USA. https://doi-org.leo.lib.unomaha.edu/10.1145/3463677.3463714.

Young, M. M., Bullock, J., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance 2*(4), 301–314.

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2021a). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4*(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006.

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly 38*(3), 1–19. https://doi.org/10.1016/j.giq.2021.101577.

CHAPTER 22

SYSTEM SAFETY AND ARTIFICIAL INTELLIGENCE

ROEL I. J. DOBBE

Introduction

The emergence of AI systems in public services, as well as in public spaces and infrastructures, has led to a plethora of new hazards, leading to accidents with fatal consequences (Raji & Dobbe, 2020) and increasing concerns over the risks posed to democratic institutions (Crawford et al., 2019). The criteria for identifying and diagnosing safety risks in complex social contexts remain unclear and contested. While various proposals have emerged specifying *what* it is we need to strive for in terms of dealing with ethical, legal, and societal implications of AI systems, there still is a long way to go to understand *how* to translate such higher-level principles and requirements to the development as well as the operation and governance of these systems in practice (Mittelstadt, 2019; Dobbe et al., 2021). A system perspective is needed (Zuiderwijk et al., 2021) that facilitates the identification of (new) complex safety hazards in ways that do justice to those bearing the brunt of AI systems (Benjamin, 2020), and can contribute to strengthening the rule of law.

While there is still considerable disagreement on what entails an appropriate definition of AI, recent policy efforts led by the OECD have converged on a seeing AI as a *sociotechnical system with a complex lifecycle* (OECD, 2021). Here we define sociotechnical AI systems as consisting of technical AI artefacts, human agents, and institutions, in which the AI artefact influences its real or virtual environment by automating, supporting, or augmenting decision making. The technical AI components in such a system may vary from logic or knowledge-based models to machine learning models to statistical or Bayesian to search or optimization methods. In systems engineering, safety is understood as an *emergent* property, which can only be instantiated and controlled for across the above-mentioned system elements (Leveson, 2012).

Core to the establishment of a system perspective are lessons from decades of knowledge built up about what constitutes safety in systems subject to software-based automation. These lessons have become central to the organization of many domains and markets, such as in aviation and healthcare. However, scanning both AI systems literature as well as policy

proposals, it is evident that these lessons have yet to be absorbed. In this chapter, I provide an on-ramp into the system safety canon for different disciplines and people involved or invested in the safeguarding of AI systems. I center the seminal work of Nancy Leveson, a pioneer of system safety in engineering systems. In her magnum opus *Engineering a Safer World: Systems Thinking Applied to Safety*, Leveson draws seven lessons based on recurring issues that hinder the safeguarding of complex systems subject to forms of software based automation (Leveson, 2012). I coin these the *Leveson Lessons* and interpret them for a variety of concrete challenges that are emerging for safeguarding AI systems in different domains.

Leveson's Lessons for AI System Safety

In this section, the seven Leveson lessons each inform a core implication for the emerging field of research and practice in AI system development and governance, and some concrete examples of approaches and tools from system safety that can form on-ramps and inspiration for *how* to put principles for safe AI systems into practice. Table 22.1 provides an overview of all lessons, implications and example system safety strategies. These tools should not be taken as a comprehensive fix for AI system safety challenges, but as a starting point to overcome gaps in mindset across policy makers, system designers, and managers, and to ensure stakeholders implicated by potential hazards have a voice in efforts to safeguard AI systems.

Shift focus from component reliability to system hazard elimination

Leveson Lesson 1: High reliability is neither necessary nor sufficient for safety.
Recent work in the area of "AI safety" has largely focused its efforts on the internal technical components and assumptions of the AI subsystem (Amodei et al., 2016). This formulation of AI systems puts most of its emphasis on the mathematical formulation, including the objective or reward function, the model class with its input variables and data, its parameters, and the task it is trying to model. As a result, it does not comprehensively address how a system is used in practice and interacts with other (human) agents, systems, and its broader environment. In a recent paper, Raji and Dobbe show that a focus on technical components alone cannot explain AI system accidents (Raji & Dobbe, 2020). They consider the use of machine learning models to predict potential criminal activity. To ensure the resulting *predictive policing system* is "fairness," various technical efforts propose metrics and tools to prevent some form of bias in the model's outputs. However, when the model is used by inherently biased police practices (Richardson et al., 2019), the discretion of police forces will still lead to more arrests in certain over-policed neighbourhoods (Lum & Isaac, 2016). Data collected from such practices can then be used to retrain the model or test new systems, leading to forms of *emergent* bias and discrimination that can't be prevented by the logical components of the AI subsystem (Dobbe et al., 2018).

Relevant system safety strategy: Identify hazards at the systems rather than component level

Rather than being able to be decomposed into components and then examined and designed for safety, AI systems are situated in complex contexts and interact with other components, both in terms of human agents, social organizations, and other technical systems. System theorists call the behavior that such systems exhibit *organized complexity* (Leveson, 2012). Typically, these systems are too complex for the sole use of "divide and conquer" analysis and design common in the traditional (engineering) sciences. In addition, such systems are not regular and random enough in their behavior to be studied purely statistically (which we return to later).

In Leveson's system safety perspective, a *hazard* is carefully defined as "[a] system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss)" (Leveson, 2012, p. 184). This definition limits the focus to states the system should never be in and gives designers greater freedom and ability to design hazards out of the system.

These fundamentals provide two important insights. First, it makes most sense to draw system boundaries in ways that include conditions related to accidents over which system designers have some control. This enables the system designer to translate identified hazards into concrete requirements, typically formulated as a constraint on design or operation of the system.

Second, system developers typically will not have control over all conditions related to accidents. Therefore, for the identification and elimination of hazards to be complete and properly assigned in terms of responsibilities, one needs to also address the institutional context. In system safety, institutional design is referred to as the *safety control structure*, which will be further discussed in this chapter. The need to integrate hazard analysis in both system design as well as operation and the broader institutional safety control structure is depicted in Figure 22.1.

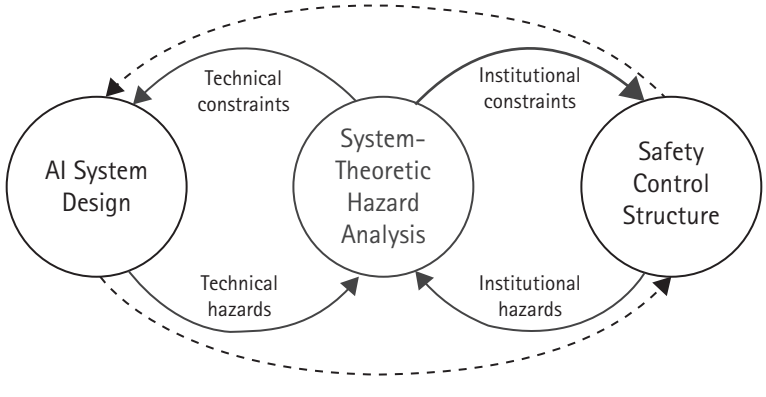

FIGURE 22.1 Hazard analysis as integrally informing the design of AI systems (in black), as well as the design of the institutional safety control structure (in blue).

Credit: Adapted by the author from Leveson (2012, chapter 9).

Let's return to the predictive policing system example. To properly identify the discriminatory nature of a predictive policing system, the system designer may draw the system boundary to include the contexts from which the data originates and in which the AI tool is used. This would naturally include the issue of institutional discrimination related to the use of these tools in the system-theoretic safety analysis and prevent naive application of these tools in practice.

Shift from event-based to constraint-based accident models

Leveson Lesson 2: Accidents are complex processes involving the entire sociotechnical system. Traditional event chain models cannot describe this process adequately.

Accidents are often studied and attempted to be explained in terms of "root causes." In doing so, many accident models rely on identifying chains of causal events. Leveson argues that a focus on causal event chains tends to narrowly isolate technical factors, engineering activities, and operator errors, thereby overlooking systemic factors that could inform prevention of future accidents.

A similar critique has surfaced recently for causal and counterfactual methods that aim to capture the "data generating process" of machine learning models, in an effort to align model-based decisions with causal structures. Hu and Kohler-Hausmann (2020) show how such approaches tend to mistake *constitutive* relationships for causal relationships in an effort to answer *why* questions. Constitutive relationships comprise variables that together constitute a category, but that do not have a clear temporal relationship, and therefore cannot be modeled as causal event-chains. Barocas et al. (2020) provide a more comprehensive account of the limitations of counterfactual methods proposed for improving the quality and fairness of decisions supported or automated by machine learning models, showing that counterfactual insights do not necessarily map to viable actions in the real-world and overlook other dimensions of the decision subject or process.

Relevant system safety strategy: Ensure safety through socio-technical constraints

Rather than trying to capture and explain safety concerns with causality, a systems-theoretic view uses *constraints* on a socio-technical system's components and their interactions, and studies and designs how and to what extent these can be effectively controlled for. Here, *control* should be interpreted to not only the role of operators or automation, but can also be attained by integrating constraints in the (physical) design of the system or through social forms of control, which may include organizational, governmental, and regulatory structures (such as rules for how a system can or cannot be used), but they may also be cultural (such as building a safe culture to report issues that could lead to safety hazards) (Leveson, 2012).

Here, we look at an autonomous vehicle example to see how system-level constraints can be identified, and how responsibility for enforcing these must be divided up and allocated

to appropriate stakeholders. On May 7, 2016, a Tesla Model S was using its "Autopilot" feature and crashed near Williston, Florida, leading to the death of the driver, Joshua Brown. The car did not brake when a tractor trailer drove across the highway perpendicularly. In its initial news release, Tesla stated that "[n]either autopilot nor the driver noticed the white side of the tractor-trailer against a brightly lit sky, so the brake was not applied," also putting heavy emphasis on the need for drivers to remain engaged with hands on the wheel while using the "Autopilot" feature (The Tesla Team, 2016).

While it acknowledged the role of human errors (both the truck driver and the Tesla driver), the National Transportation Safety Board (NTSB) determined that "the operational design of the Tesla's vehicle automation permitted the car driver's overreliance on the automation, noting its design allowed prolonged disengagement from the driving task and enabled the driver to use it in ways inconsistent with manufacturer guidance and warnings" (The National Transportation Safety Board, 2017, p. 42). In addition, it found that "[t]he Tesla's automated vehicle control system was not designed to, and could not, identify the truck crossing the Tesla's path or recognize the impending crash" (p. 30).

While this verdict does not directly specify what safety constraints were missing or unsuccessfully enforced, it does provide suggestions. Clearly, if the autonomous driving feature continues, there will need to be put in place better conditions to prevent prolonged disengagement and unsafe use. The faulty vision system needs a structural reconceptualization to either ensure that predictable scenarios, like a truck crossing, are detectable or to constrain the use of the vision system only to situations where its errors cannot lead to hazardous conditions.

Shift from a probabilistic to a system-theoretic safety perspective

Leveson Lesson 3: Risk and safety may be best understood and communicated in ways other than probabilistic risk analysis.

Most AI systems parameters are optimized according to some objective or reward function. These tend to either minimize or maximize a probability that the AI model output matches a desirable outcome. It is no secret that such approaches will never *guarantee* that an AI system's output enters unsafe territory—they may only reduce the probability. Just as probabilistic risk assessments, the reward or objective function tends to absorb a plethora of risky dynamic safety scenarios into one function without explicitly formulating these scenarios, let alone how to safeguard against these.

Internalizing constraints in the design and training of ML-driven AI systems is notoriously hard and, while some relevant efforts have been made in theory (Achiam et al., 2017), a workable and scalable solutions is not yet available.

Solely relying on an AI system's internal model to incorporate all safety constraints is dangerous. Raji and Dobbe (2020) note that you can expect AI system models to fail and hence, of true scrutiny should be the effectiveness of the *safety measures* employed, such as fail-safe features to kick in when an anticipated engineering failure occurs. Instead of capturing safety in terms of low probabilities, alternative methods are needed to secure

a system in practice, by combining learning-methods with control-theoretic guarantees (Fisac et al., 2019) and installing other (socio-technical) fail-safe mechanisms (Dobbe et al., 2021). To design those effectively, we first must reevaluate where we draw the boundary of our system design and accompanying safety analysis.

Relevant system safety strategy: Capture the safety conditions and assumptions in a process model

The boundary of analysis for safety should extend beyond the AI model (with its inputs, features, outputs, and reward/objective function), and include dynamics of the AI model interacting with other system components, including human operators, users, or subjects, and the process and environment the system is intervening in. In system safety, this is done with a *process model*, which interprets an AI system as a control entity, possibly complemented by forms of human control and decision-making, as indicated in Figure 22.2. The process model formulates four conditions:

1. *The goal:* the objectives and safety constraints that must be met and enforced by the controller;
2. *The action condition:* the controller must be able to affect the state of the system;
3. *The observability condition:* the controller must be able to ascertain the state of the system, through feedback, observations, and measurements; and
4. *The model condition:* the controller must be or contain a model of the process. A human controller should also have a model of the behavior of the AI techniques used for control and decision-making.

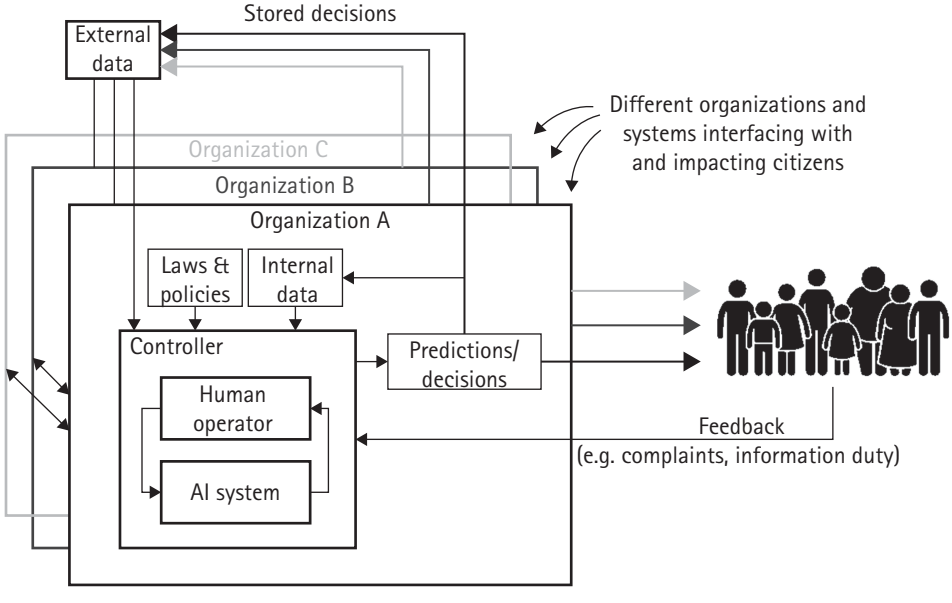

FIGURE 22.2 A specific depiction of the process model in the context of several government organizations making AI-informed decisions about households that are dynamically linked through various data registers.

Credit: Author.

Based on the process model, it is easier to specify what feedback and fail-safe mechanisms are needed to enforce safety constraints, and, as a consequence, to understand how accidents occur.

Leveson argues that often, accidents occur "when the process model used by the controller (automated or human) does not match the process" (Leveson, 2012, p. 88). This can lead to four types of control errors: incorrect or unsafe control actions, missing control actions for guaranteeing safety, wrong timing of control actions, or control actions applied too long or stopped too soon.

Here we end with an example from the public sector in The Netherlands. The Dutch government developed the System Risk Indication (SyRI); a legal instrument to detect various forms of fraud, including social benefits, allowances, and taxes fraud, by developing risk profiles using statistical learning techniques. These risk profiles determined the deployment of resources to combat fraud. In 2020, The Hague District Court ruled the SyRI system and its accompanying legislation unlawful as it violated the European Convention on Human Rights (ECHR): "The court finds that the SyRI legislation in no way provides information on the factual data that can demonstrate the presence of a certain circumstance, in other words which objective factual data can justifiably lead to the conclusion that there is an increased risk" (Rechtbank Den Haag, 2020, article 6.87). Later that year, the Dutch Data Protection Agency concluded that the risk classification system had been used in the childcare benefits scandal, in which more than 26,000 families were wrongfully accused of fraud (Autoriteit Persoonsgegevens, 2020). Many families collected steep amounts of debt and were forced to foreclose and sell their homes or vehicles, also leading to massive emotional harms, including forced divorces and even cases of suicide (Levie, 2021).

While the scandal is still under investigation, it is clear that AI systems played a central role in causing these harms, with algorithms discriminating based on ethnicity and income level. This case exemplifies the sheer lack of a process model. The system informed decisions (action condition) based on illegitimate inputs (observation condition) and had no idea of the detrimental consequences for citizens (model condition), nor did it have any safeguards in place (goal condition). The AI system used was part of a broader policy and governance structure that dehumanized thousands of families. Arguably, the model of the controlled process should have included the impact of possible control errors (in this case the incorrect classification of families as "fraudulent") on the financial and general health of the subject families. That would have enabled the formulation of safety constraints that would rule out the unjustified condemnation. In the Dutch legal system, this would have included proper avenues for legal protection and due process in the case of disagreement on the verdict of the system and its operators. Such safety hazards would have been naturally identified if human dimensions were part of the model of the controlled process, which is now a top priority for many policy domains. The Dutch case shows that their public administration, while being leading in terms of digitalization, does not have a long history of dealing with safety hazards and centering (possible) victims of control errors in improving automated decision-making. As such, it will be able to learn significantly from the lessons learned in system safety.

Shift from siloed design and operation to aligning mental models

Leveson Lesson 4: Operator error is a product of the environment in which it occurs. To reduce operator "error" we must change the environment in which the operator works.

While investigations into accidents tend to structurally find other factors, there is a recurring tendency to put the blame on inadequate response to the failure by an operator (Leveson, 2012). This phenomenon has been attributed to the effect of *hindsight bias* in which the focus is on what an operator did wrong rather than on what was the logical thing to do given the circumstances (Dekker, 2016). In the context of AI systems research, the dominant focus on technical solutions often contributes to an under-appreciation for the role of human operators and other users of a system (Dobbe et al., 2021).

Meanwhile, new regulatory proposals are putting more emphasis on the need for human oversight, as a response to emerging evidence on the dangers of AI systems. The European Commission (2021) highlights the role of meaningful human oversight in their recent proposal for regulating AI systems. Kak and Green (2021) point out that this may not address societal concerns, showing how earlier calls have provided shallow protection in the past and may instead serve as a means of "rubber stamping" intrusive and harmful applications, thereby blurring responsibility, where frontline human operators of AI systems are blamed for broader system failures over which they have little or no control (Elish, 2019).

Relevant system safety strategy: Align mental models across design, operation and affected stakeholders

While issues around the operator and human–machine interactions occurring with new AI systems seem new, these have been studied extensively for safety-critical systems such as airplane autopilots. Here I focus on the importance of *mental models*, which are maintained by designers and operators and "used to determine what control actions are needed, and it is updated through various forms of feedback" (Leveson, 2012, p. 87). Mental models are a concept that finds its roots in cognitive science, engineering, and the study of work performance (Rasmussen, 1987).

Across the mental models formed by designers and operators natural changes emerge over time. Operators or users typically find rational grounds to deviate from the *normative procedures* described by designers, leading to *effective procedures* (Rasmussen et al., 1994). This *adaptation* is common across all systems and may happen due to various reasons, such as time constraints, economic or political incentives, or an operator understanding how to better safeguard or optimize a system through first-hand experience, which a designer typically lacks. It is therefore no surprise that operator error is found to be a dominant focus for accident causes, as any deviation from normative procedure could be labeled as erroneous. In turn, this effect is often used as an invalid argument to automate the operator away (Dekker, 2017).

Instead, the system safety discipline would describe efforts to align and periodically update the different *mental models* that different actors hold and use when designing, operating, or otherwise interacting with the system. In Figure 22.3, I show an adaptation of Leveson's depiction of the relationships between mental models (straight arrows). In addition to designers and operators, the increasing dominance of automation in societal infrastructures also leads to other stakeholders forming mental models of a system's behavior and how to depend on and interact with it. The updating of mental models through feedback mechanisms will be further addressed later in this chapter.

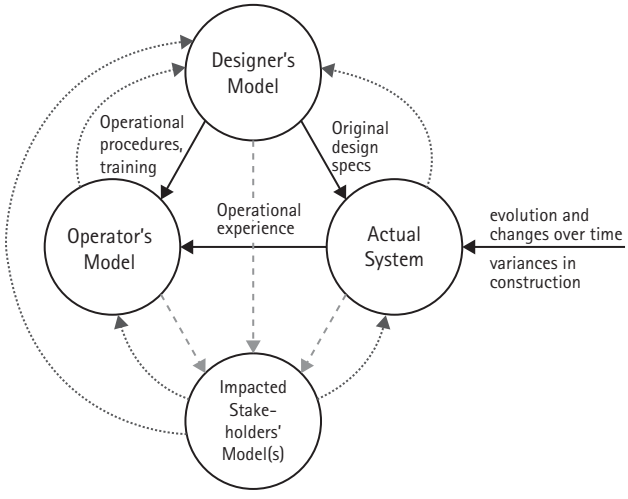

FIGURE 22.3 Relationships between mental models.

Credit: Adapted by the author from Leveson (2012).

Here we discuss the importance of mental models for a recent case in autonomous driving. The fatal accident in which an autonomous vehicle, developed and tested by Uber, killed a road-crossing pedestrian in Tempe, Arizona, on March 18, 2018, forms a painful example. When evidence emerged that the test driver, responsible for taking over control of the car in the event of unforeseen and risky conditions, had been visually distracted due to looking at her cellphone, most public journalism focused on her role in causing the accident. But while the NTSB did identify the operator's role as the probable cause of the accident, it also put major emphasis on the role Uber had played in enabling such unsafe behavior: "Contributing to the crash were the Uber Advanced Technologies Group's (1) inadequate safety risk assessment procedures, (2) ineffective oversight of vehicle operators, and (3) lack of adequate mechanisms for addressing operators' automation complacency—all a consequence of its inadequate safety culture" (The National Transportation Safety Board, 2019, p. v). The damning NTSB verdict makes clear the essence of aligning the mental models across design and operation.

While it is evident that the behavior the test driver exhibited was unsafe, there were various environmental factors Uber could have addressed to prevent this from happening. A typical reflex to this is to install human factor checklists for designers and put in place stricter procedures for operators to follow. These, however, have been shown not to be sufficient in establishing high levels of safety, and can even have the opposite effect (Dekker, 2017). Instead, engineers and designers need to design the system and the environment to prevent human errors. A general approach to designing adequate shared control of humans and AI systems is unrealistic because there are a plethora of contextual considerations (Leveson, 2012). Nevertheless, Leveson has outlined three principles:

1. *Design for redundant paths:* to provide multiple paths to ensure that a single error cannot prevent the operator from taking action to maintain a safe system state and avoid hazards;

2. *Design for incremental control:* to give the operator enough time and feedback to perform control actions and, if possible, to do so in incremental steps rather than in one control action; and
3. *Design for error tolerance:* to make sure that reversible errors are observable to the human operator before unacceptable consequences occur, to allow them to monitor their own performance and to recover from erroneous actions.

Curb the curse of flexibility in AI software development

Leveson Lesson 5: Highly reliable software is not necessarily safe. Increasing software reliability or reducing implementation errors will have little impact on safety.

Perhaps the most pertinent and under-appreciated lesson from Leveson's system-safety canon relates to the role of software in algorithmic harms. The *curse of flexibility* refers to the dominant assumption that, in contrast to physical machines, software is not subject to physical constraints, leading to ever more flexible and complex designs aiming to address existing and emerging requirements. However, Leveson argues that the limiting factors of software-based automation change from the structural integrity of materials to limits on people's intellectual capabilities (Leveson, 1995, 2012).

The source of most serious problems with software relate to outsourcing software development. This creates extra communication steps between those understanding the physical realization of the system (domain experts, users, affected stakeholders, hardware operators) and those programming the software. And this step often leads to requirements not being complete. Leveson argues that "[n]early all the serious accidents in which software has been involved in the past twenty years can be traced to requirements flaws, not coding errors. . . . The most serious problems arise, however, when nobody understands what the software should do or even what it should not do" (Leveson, 2012, p. 49). Unfortunately, there is mounting evidence that the development of high-stakes AI systems and software often goes without principled ways to determine safety requirements and translate these to software implementations (Dobbe et al., 2021).

The resulting lack of proper requirements and validation is best reflected in the development of large language models (LLMs) within the field of natural language processing (NLP). LLMs are notorious for their complexity and flexibility. This combined with the struggle to impose reasonable safety requirements make these systems unsafe by design, with large environmental consequences (Dobbe & Whittaker, 2019), damage to cultural heritage and harm to groups and individuals subject to an LLM's errors (Bender et al., 2021).

Relevant system safety strategy: Include software and related organizational and infrastructural dependencies in system-theoretic hazard analysis

The System-Theoretic Process Analysis (STPA) that Leveson developed formalized hazard analysis to provide information and documentation necessary to ensure the enforcement of safety constraints in system design, development, manufacturing, and operations (Leveson, 2012). The output of such analysis can be used to develop or update the

requirements of a software-based automation system in the process model, as well as the institutional arrangements in the safety control structure.

STPA considers software as a primary source for possible hazards, as it modulates decisions and control actions, provides feedback (e.g., through sensor readings) and executes computational models. A more nascent application of STPA is to use it to identify safety risks related to how the software is organized. Over the last decade, we have seen a major shift to *cloud computing platforms*, in which a broad variety of services for software development and maintenance are integrated, including tools for integrating data and building AI applications. Concurrently, AI applications have been championed to solve complex problems in marketing efforts by vendors, both those developing cloud platforms and third parties working with it. The AI Now Institute showed how the AI hype has contributed to "widening gaps between marketing promises and actual product performance. With these gaps come increasing risks to both individuals and commercial customers, often with grave consequences" (Whittaker et al., 2018, p. 5).

Gurses and van Hoboken (2017) show how cloud computing platforms and their central bundling of services have challenged principled software development, rendering designing for crucial values like safety or privacy in AI system software an elusive goal. More alarmingly, the control of big tech companies over infrastructures central to development of high-risk and high-impact AI systems threatens our ability to understand and regulate these technologies (Whittaker, 2021).

As a result, there is a set of open problems related to safeguarding the organizational and infrastructural risks of software used to build AI systems in critical public domains, in ways that overcome and rebalance power relations (Balayn & Gürses, 2021). Here, a system safety can help us to expand the boundary of safety analyses to include organizational and infrastructural dependencies, to directly address the hazards that emerge through contracts that lean on central bundling of software services, and to characterize the risks that emerge from not being able to program safety requirements in an end-to-end fashion.

Translate safety constraints to the design and operation of the system

Leveson Lesson 6: Systems will tend to migrate toward states of higher risk. Such migration is predictable and can be prevented by appropriate system design or detected during operations using leading indicators of increasing risk.

Through a safety-guided design, many hazards may be anticipated and eliminated. However, there are various reasons why hazards may arise during operations. After decades of research, system safety pioneer Jens Rasmussen concluded that systems tend to migrate toward states of higher risk, and that such adaptation is not random but predictable and, therefore, manageable (Rasmussen, 1997). An important recurring observation in his research shows that safety defenses are likely to degenerate, especially "when pressure toward cost-effectiveness is dominating," and that often accident investigations conclude that the particular accident was "waiting for its release" (p. 189). The *safety control structure* should hence assign responsibilities for monitoring and overseeing that the system operates safely subject to inherent adaptation.

Relevant system safety strategy: Organize feedback mechanisms for operational safety

An operational safety control structure establishes controls and feedback loops to "(1) identify and handle flaws in the original hazard analysis and system design and (2) to detect unsafe changes in the system during operations before the changes lead to losses" (Leveson, 2012, p. 394). Here I focus on the value of three concrete feedback mechanisms for maintaining and improving operational safety, namely audits, accident investigations, and reporting systems. Note that these mechanisms are different from the feedback that the controller (a combination of the AI system and/or a human operator) receives in real-time to control a process (as depicted in Figure 22.2). Instead, this feedback forms input to reevaluate the system design, to update safety constraints (based on hazard analysis), and to update the process model and roles and responsibilities in the safety control structure.

Audits and performance assessments are done to determine whether safety constraints are enforced in the operation of the system, and whether the rationale of the system design holds up in practice. The need for auditing AI systems is still in its infancy, but there have been some efforts to audit for the presence of bias (Raji & Buolamwini, 2019). Effectuating functional audits may be a struggle because behaviors may change in anticipation of an audit or a lack of independence may influence its objectivity. Overcoming such issues requires building a conducive organizational culture. A *participatory audit* aims to impact the cultural challenges (Leveson, 2012, chapters 12 and 14), partly by not making audits punitive and ensuring all levels of the safety control structure are audited, including affected stakeholders. This aspect is corroborated by Rasmussen's findings, who stresses the need "to represent the control structure involving all levels of society for each particular hazard category" (Rasmussen, 1997, p. 183).

Accident, incident, and anomaly investigations are meant to draw lessons and potential conclusions for improving the system design, process model, and safety control structure. These investigations have historically suffered from *root cause seduction*, meaning a too narrow focus on finding root causes to explain complex accidents (Carroll, 1995). Instead, as previously discussed, investigations should use a systems perspective for understanding accidents. A crucial component is to develop such a perspective in the organizations so that investigations can be done holistically and lead to follow-up and training of all involved professionals.

Finally, reporting systems can help to detect issues before they turn into hazards. Reporting systems are common in safety-critical domains, such as healthcare or aviation. As discussed for the other feedback mechanisms, there are hurdles to ensuring a reporting system is used. It is therefore important to understand why people won't use them and fix those. Common factors have to do with an ineffective or unclear interface of the system, a lack of follow-up, or fear that reported information may be used against the person or someone else.

Build an organization and culture that is open to understanding and learning

Leveson Lesson 7: Blame is the enemy of safety. Focus should be on understanding how the system behavior as a whole contributed to the loss and not on who or what to blame for it.

As described earlier, many major accidents in complex systems are met with subjectivity. Safety is a deeply normative concept that is understood, formalized, and experienced in different ways across different cultures, communities, and organizations. This subjectivity is a major source of challenges in the emerging research and practices for safeguarding AI systems (Dobbe et al., 2021). For example, the attribution of accident causes can be subject to the characteristics of the victim and the analyst, such as hierarchical level, job satisfaction, and level involvement in the accident (Leplat, 1984).

The emerging harms related to AI systems have motivated a plethora of ethical principles, guidelines, and policy instruments (Schiff et al., 2021). While safety is a core tenet of recent regulatory efforts (European Commission, 2021), the accountability gaps existing for many AI-driven innovations (Whittaker et al., 2018) may be a source of blame games and finger pointing that can take away attention from understanding how accidents may happen or have happened as a way to do better at building and operating AI systems. While sometimes misinterpreted as a means to divert accountability from humans to the system, the establishment of system safety has contributed greatly to new more effective ways to promote and structure individual responsibility and accountability (Dekker & Leveson, 2015), especially when combined with cultural change programs that promote sharing of information to improve safety, as evidenced in the healthcare and aviation industry (Dekker, 2016).

Relevant system safety strategy: Balancing safety and accountability through a just culture

Leveson's last lesson is the most crucial. In the final chapters of her book *Engineering a Safer World: Systems Thinking Applied to Safety*, Leveson underlines the importance of adequate management and safety culture to accomplish any of the goals previously described (Leveson, 2012). The increasing scale and complexity of AI systems makes that these challenges span broader institutional networks, often comprising public, private, knowledge, and societal institutions (Janssen & Kuk, 2016).

Leveson outlines the core organizational requirements managers need to meet for improving safety: an effective safety control structure, a safety information system, and a strong and sustainable safety culture. Here we focus on the often underappreciated importance of the latter.

Bundled in his book *Just Culture: Balancing Safety and Accountability*, Sidney Dekker describes how to stray away from detrimental blame cultures, instead arguing for the crucial need to provide a comfortable environment in which people feel safe to share information about what should be improved to those who can do something about it, and to allow for investments in improvements and learning that have a real safety dividend rather than spending money on legal protection and limiting liability (Dekker, 2016). Through his extensive field work and experience as a pilot, Dekker explains the struggles that organizations go through to build an effective safety culture. Unfortunately, there is no simple formula, but Dekker has gained important insights into *how* to establish such cultures and core agreements on system safety procedures and responsibilities: "What matters for organizational justice is not so much whose version gets to rule the roost—but

to what extent this is made explicit and how it is decided upon" (Dekker, 2016, 75). The key steps he outlines are:

1. Design the process to deal with adverse events or apparently risky acts, which does not interfere with performance review and makes explicit how judgments are made and enables opportunities for appeal.
2. Decide who is involved in this process, in a way that allows for impartial input from across the organization and includes domain expertise to represent and acknowledge work floor complexity.
3. Decide who is involved in deciding who is involved, to prevent top-down decision-making that lacks buy-in and ownership.

The aviation industry is historically known for its safety culture, as also studied by Dekker (Dekker, 2016). In recent years, two crashes occurred with a Boeing 737-MAX, in which the plane took an uncontrollable nosedive. It turned out that something had gone wrong with the aircraft calibration algorithm (Gelles, 2020). Driven by competitive forces to update the 737 models with new engine technology, the company decided to bypass certain checks on how new engines would affect the autopilot behavior, ultimately causing pilots to be confused and unable to steer the plane out of a nosedive. As such, despite the broader industry's emphasis on safety culture, Boeing's own culture had been a breeding ground for irresponsible decisions and unsafe design choices (Boeing, 2019). This example shows the detrimental effects that organizational culture and decision-making can have on engineering choice, and in effect the safety of a system in operation. Hence, organizations responsible for such systems should build a strong safety culture and have the leadership and mechanisms in place to maintain it.

Conclusions

In this chapter, I covered central insights from the field of system safety, with particular focus on what it takes to safeguard systems that are subject to software-based automation. The lessons and concrete system safety strategies are summarized in Table 22.1. We have seen that most of these lessons have yet to be absorbed in domains where AI systems are emerging to inform, mediate, or automate decision-making.

It would, however, be naive to conclude that the strategies presented here can be straightforwardly applied. AI systems' inherent flexibility and associated failure modes make designing for safety, as well as formulating effective process models and safety control structures a complex task. If anything, the power of evidence that system safety strategies have could help understand when using an AI system is a good idea to begin with. Formulating constraints for valid safety problems can bring sanity to what is "responsible AI" by making explicit when and how not to use it, to prevent unnecessary harm.

Lastly, it is crucial to note that any tool or strategy can be applied in irresponsible ways, system safety approaches included. It should be evident from this chapter that applying these strategies requires various forms of formalization and modeling. As such, system safety methods will always redraw, solidify, and impose power relationships. This is not necessarily a

Table 22.1 Overview of Leveson lessons, implications for AI system development and governance and examples of relevant system safety strategies.

Leveson Lesson	AI System Safety Implication	Example System Safety Strategy
Component reliability is insufficient for safety	Identify and eliminate hazards at system level	System hazard-informed system design and safety control structure
Causal event models cannot capture system complexity	Understand safety through sociotechnical constraints	System-theoretic accident models: integrating safety constraints, the process model, and the safety control structure
Probabilistic methods do not provide safety guarantees	Capture safety conditions and requirements in a system-theoretic way	Process model: AI system goals, actions, observation, and model of controlled process and automation
Operator error is a product of the environment	Align mental models across design, operation, and affected stakeholders	Leveson's design principles for shared human–AI controller design: redundancy, incremental control, and error tolerance
Reliable software is not necessarily safe	Include (AI) software and its organizational dependencies in hazard analysis	System-theoretic process analysis
Systems migrate to states of higher risk	Ensure operational safety	Feedback mechanisms (audits, investigations, and reporting systems)
Blame is the enemy of safety	Build an organization and culture that is open to understanding and learning	Just Culture

bad thing, but it should be made explicit in an effort to empower those that need safeguarding. Just as patient safety has become the central focus in successful healthcare systems, vulnerable people and communities should have a seat at the table and see themselves empowered in the crystallization of a system's design and operational safety control structure.

References

Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70* (pp. 22–31). ICML'17, Sydney, NSW, Australia.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Man´e, D. (2016). Concrete problems in AI safety. arXiv:1606.06565 [cs].

Autoriteit Persoonsgegevens. (2020). *De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag*. Autoriteit Persoonsgegevens.

Balayn, A., & Gu¨rses, S. (2021). *Beyond Debiasing—Regulating AI and Its Inequalities*. Technical report, EDRi.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 80–89). FAT* '20. New York, NY, USA. Association for Computing Machinery.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). FAccT '21. New York, NY, USA. Association for Computing Machinery.

Benjamin, R. (2020). Race after technology: Abolitionist tools for the new Jim code. *Social Forces 98* (4), 1–3.

Boeing. (2019). Boeing Chairman, President and CEO Dennis Muilenburg Announces Changes to Sharpen Company Focus on Product and Services Safety, September 30.

Carroll, J. S. (1995). Incident reviews in high-hazard industries: Sense making and learning under ambiguity and accountability. *Industrial & Environmental Crisis Quarterly 9* (2), 175–197.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., & Sa´nchez, A. N. (2019). *AI Now Report 2019*. AI Now Institute.

Dekker, S. (2016). *Just culture: Balancing safety and accountability*. CRC Press.

Dekker, S. (2017). *The field guide to understanding "human error"*. 3rd ed. CRC Press.

Dekker, S. W. A., & Leveson, N. G. (2015). The systems approach to medicine: controversy and misconceptions. *BMJ Quality & Safety 24* (1), 7–9.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. In arXiv preprint arXiv:1807.00553. Presented at the Workshop on Fairness, Accountability and Transparency in Machine Learning, 2018 International Conference on Machine Learning in Stockholm.

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence 300*, 103555.

Dobbe, R., & Whittaker, M. (2019). *AI and Climate Change: How they're connected, and what we can do about it*. Technical report, AI Now Institute.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human–robot interaction Engaging Science, Technology, and Society, April 3. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757236.

European Commission (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | Shaping Europe's digital future. Technical report, European Commission, Brussels.

Fisac, J. F., Akametalu, A. K., Zeilinger, M. N., Kaynama, S., Gillula, J., & Tomlin, C. J. (2019). A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems. *IEEE Transactions on Automatic Control 64* (7), 2737–2752.

Gelles, D. (2020). "I honestly don't trust many people at Boeing': A broken culture exposed. *The New York Times*, January 10.

Gurses, S., & van Hoboken, J. (2017). Privacy after the agile turn. *SocArXiv*. https://osf.io/preprints/socarxiv/9gy73/.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with fair machine learning? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (p. 513. Barcelona.

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly 33* (3), 371–377.

Kak, A., & Green, B. (2021). The false comfort of human oversight as an antidote to AI harm. *Slate Magazine*, June 15. https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html.

Leplat, J. (1984). Occupational accident research and systems approach. *Journal of Occupational Accidents 6* (13), 77–89.

Leveson, N. G. (1995). *Safeware: System safety and computers*. Addison-Wesley.

Leveson, N. G. (2012). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.

Levie, K. (2021). The Dutch Government's Benefits Scandal Is Rooted in Stigma Against Welfare Recipients. *Jacobin Magazine*, January 23. https://jacobinmag.com/2021/01/dutch-welfare-benefits-childcare-scandal.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance 13* (5), 14–19.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1* (11), 501–507.

OECD. (2021). *Recommendation of the council on artificial intelligence*, OECD/LEGAL/0449. Technical report, OECD.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429–435). AIES '19, New York, NY, USA. Association for Computing Machinery.

Raji, I. D., & Dobbe, R. (2020). Concrete Problems in AI Safety, Revisited. In *Workshop on Machine Learning In Real Life at the International Conference on Learning Representations*. Addis Abeba.

Rasmussen, J. (1987). *Mental models and the control of actions in complex environments*. Technical Report 2656, Risø National Laboratory.

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science 27* (2–3), 183–213.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. 1st ed. Wiley-Interscience.

Rechtbank Den Haag. (2020). *ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag, C-09-550982-HA ZA 18-388 (English)*. Uitspraak.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review 94*, 192–233.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society 2* (1), 31–42.

The National Transportation Safety Board. (2017). Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida May 7, 2016. Accident Report NTSB/HAR17/02 PB2017-102600. The National Transportation Safety Board, Washington, DC.

The National Transportation Safety Board. (2019). Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018. Accident Report NTSB/HAR-19/03 PB2019101402. National Transportation Safety Board, Washington, DC.

The Tesla Team. (2016). A Tragic Loss. *Tesla Blog*, June 30. https://www.tesla.com/blog/tragic-loss.

Whittaker, M. (2021). The steep cost of capture. *Interactions 28* (6), 50–55.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now report 2018*. AI Now Institute at New York University.

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly 38* (3), 101577.

SECTION V

ASSESSMENT AND IMPLEMENTATION OF AI GOVERNANCE

MATTHEW M. YOUNG AND YU-CHE CHEN

CHAPTER 23

ASSESSING AUTOMATED ADMINISTRATION

CARY COGLIANESE AND ALICIA LAI

Public administration is susceptible to a series of well-documented biases and sources of error stemming from both human cognition and group decision-making. At times, these limitations of human judgment have unfortunately led to major tragedies and troublesome inefficiencies. To overcome the limitations of human decision-making, while concurrently responding to pressing governance needs in the face of limited resources, public administrators are beginning to explore new possibilities offered by artificial intelligence (AI) (Coglianese & Ben Dor, 2021; Engstrom et al., 2020).*

AI tools promise to increase the accuracy, consistency, and speed of governmental decisions and the performance of key tasks. But these digital tools also raise their own set of legal, management, and political issues. When, if at all, should government agencies use automated AI tools to substitute for existing human-based processes?

This chapter assesses this question, offering a framework for deciding whether and when administrative agencies should use AI tools that replace or significantly augment human decision-making. These tools are themselves not perfect, but the relevant question confronting public administrators is whether they do better than human decision-making at specific tasks. Assessing the promise of automated administration will ultimately depend on making concrete, relative assessments of whether specific AI tools can improve the status quo in specific contexts. Nevertheless, it is possible to offer some general considerations to account for when making specific assessments.

We begin by presenting a general case for using automated AI tools in the public sector. We next discuss the core legal issues that are likely implicated by governmental reliance on AI tools. These tools can and do raise a range of legal issues, but we conclude that there currently exist no intrinsic legal barriers to the use of AI, at least under prevailing administrative law doctrines in the United States. As long as public administrators approach the development and deployment of these tools with due care and responsibility, the use of AI can comfortably fit within existing legal doctrine. We conclude by offering guidance for the ultimate challenge for public administrators: deciding when these tools should replace human decision-making in the performance of specific tasks, and then determining how to design and implement these tools in a responsible manner.

The Case for Automating Administration

Governmental tasks today depend largely on human decision-making. As a result, many of government's failings—from major tragedies to minor inefficiencies—can be traced to limitations on human decision-making. Even when humans act with the best of intentions, their decisions can be prone to a wide array of well-documented cognitive biases, physical limitations, external pressures, and basic missteps. Additional dysfunctionalities come into play when groups of humans make collective decisions, as they often do in government.

Consider, for example, a few cases of major governmental breakdowns that have stemmed in substantial part from limitations in individual and collective decision-making:

- When forced to confront a pandemic in early 2020, the U.S. federal government found itself ill-prepared and fumbling in its response due to a variety of cognitive, bureaucratic, and political factors. For years prior, governmental leaders tended to discount the risks of and under-prepare for a major pandemic in the United States (Fisher, 2020; Lewis, 2021; Slavitt, 2021). Even after the outbreak emerged, government officials and many Americans found themselves prone to downplaying risks and resisting key public health actions, such as social distancing, masking, and vaccinations (Halpern et al., 2020).
- The 1986 Space Shuttle Challenger exploded on liftoff, resulting in the deaths of all seven crew members, due to a series of individual and collective dysfunctions in deciding whether to launch the shuttle in cold weather (Report of the Presidential Commission, 1986; Vaughan, 2016). Group pressures among NASA managers and Thiokol representatives played a pivotal role in the decision, as did the fatigue that had set in for the key decision-makers (Forrest, 2005).
- With the Bay of Pigs mission in 1961, human and organizational factors led to over one hundred deaths in the failed attempt by the United States to overthrow the Castro government in Cuba (Wyden, 1979). Key decision-makers had succumbed to a sunk-cost fallacy. Moreover, President Kennedy and his advisors asked operational leaders few questions about risks and back-up options and they failed to invite alternative perspectives (Neustadt & May, 1986).

In raising these debacles from across the decades, we do not necessarily claim that AI would have prevented them but instead we offer these examples to show how the shortcomings of human decision-making can play a significant role in affecting public administration.

Unfortunately, these examples are not nearly as exceptional as they are prominent (Light, 2014; Schuck, 2014). To these examples, we can add a list of delays, biases, errors, and inconsistencies that afflict other, more routine decisions made by individual actors within government. A legion of frailties and foibles show that human decision-making is far from perfect. At an individual level, the range of shortcomings can be grouped into at least six categories:

1. *Memory, fatigue, and aging*. Human working memory is limited in its capacity (Cowan, 2001). Fatigued individuals are less alert and slower to react, and they

experience lapses in memory and generally have difficulty processing information (Alhola, 2007). Over time, aging causes the human brain to shrink in size and its memory to decline (Peters, 2006).

2. *System neglect*. People can overemphasize signals relative to the underlying system which generates the signals, such as in financial sectors (Kremer et al., 2011).
3. *Loss aversion*. People tend to dislike losses far more than they like corresponding gains of the same magnitude (Freund & Özden, 2008).
4. *Hindsight and availability bias*. People increase their estimates of probabilities of past events when retroactively considering an event (Fischhoff, 1975). Recall is subject to availability bias, giving prominence to hazards that are more cognitively available through recent and memorable instances of harm (Tversky & Kahneman, 1974).
5. *Confirmation bias*. People tend to search for and favor information that confirms their existing beliefs, and they tend to ignore or discount information that is inconsistent with or challenges those beliefs (Lord et al., 1979).
6. *Racial and gender biases*. People are subject to implicit biases in making judgments whenever race, gender, and other similar characteristics are involved (Pascalis, 2005; Eberhardt, 2019).

These various limitations of individual judgment ultimately can affect governmental decision-making on a regular basis. For example, judges' criminal sentencing decisions have been shown to vary systematically based on the race of the defendant (Rehavi & Starr, 2014). Human processing of social security disability benefits produces results that differ markedly from one administrative official to another (TRAC, 2011). Elected politicians tend to process information selectively, emphasizing the information that confirms their prior beliefs (Baekgaard et al., 2019).

Although it may often be thought that group decisions bring added wisdom once people can "put their heads together," it is far from clear that group decision-making always fares better than individual decision-making. Research generally shows that "[i]f some bias, error, or tendency predisposes individuals to process information in a particular way, then groups exaggerate this tendency" (Hinsz et al., 1997). Moreover, collective decision-making brings its own distinctive pathologies and limitations. One of these pathologies, known as groupthink, follows from a psychological drive for consensus that suppresses dissent and the appraisal of alternatives (Janis, 1972). Another related pathology, social loafing, stems from reductions in motivation and accountability (Kogan & Wallach, 1967), and it manifests in slacking off by group members (Latane et al., 1979). Social loafing also corresponds with the well-known problem of collective action, in which individuals have an incentive to free-ride on the production of any collective good (Olson, 1965). In addition, even if all members of a group are active and well-motivated, their collective decisions may lead to incoherent outcomes due to the cycling of preferences among the group's different members (Arrow, 1950). It was not without reason that Otto von Bismark quipped that the making of laws reminded him of the making of sausages. Some research indicates that half of all group decisions made within organizations end up in failure (Nutt, 1999).

The many limitations of human decision-making—both individual and collective—almost singlehandedly make the case for governmental use of AI tools. Although automated alternatives to human judgment cannot eliminate every human error, AI tools can be designed to "perform functions that are normally associated with human intelligence

such as reasoning, learning, and self-improvement" (NIST, 2021). They offer the prospect for improving the accuracy and consistency of a variety of governmental tasks, as well as reducing delays and improving administrative efficiencies (Lai, 2021; Sunstein et al., 2021).

Studies have shown that algorithms can more accurately recall memorized content than humans can (Panigrahi, 2018). They can also do better than humans in terms of predictive accuracy (Agrawal et al., 2018). Research indicates that, if applied to bail decisions, machine-learning algorithms could be expected to reduce crime rates substantially even with no change in the rate of jailing—or, alternatively, they could reduce jailing rates up to 42 percent with no change in crime rates (Kleinberg et al., 2018). If used by environmental regulators to select industrial facilities to inspect for compliance with water pollution regulations, machine-learning algorithms could help increase the identification of violators by as much as 600 percent (Hino et al., 2018).

Automated systems may also promote greater uniformity of decision-making, especially for tasks that currently necessitate repeated judgments involving multiple individuals, such as decisions concerning taxation, immigration status, or benefits eligibility. The alternative to automation—namely, the training and overseeing of human decision-makers—can be costly or ineffective (or both). But a uniform algorithmic system that applies nationwide may feasibly achieve more consistent results across individual cases (Shrestha, 2019). Furthermore, when errors arise or biases are discovered, automated systems that rely on a common algorithm may be easier to modify and fix than to re-train a large number of human decision-makers.

In an important sense, the case for AI is already being made by its widespread adoption and use in the private sector (McKinsey Global Institute, 2018). Digital algorithms are being shown to outperform experts in tasks as varied as medical diagnosis, making mortgage lending decisions, and placing bets on sporting events (Tschandl, 2019; Gates, 2010; Pretorius & Parry, 2016).

Public administrators are also realizing that AI holds promise. Federal, state, and local agencies have begun using automated tools for a variety of administrative tasks (Engstrom et al., 2020; Bray, 2014). The Bureau of Labor Statistics, for example, currently uses an AI system to categorize reports of workplace injuries submitted by over 200,000 businesses (Chenok & Yusti, 2018; BLS, 2019). The Food and Drug Administration uses AI for real-time tracking and reporting of microbial sources in foodborne outbreaks (FDA, 2011). The Federal Communications Commission has used an AI tool to sort and analyze the 22 million public comments it received in connection with its net neutrality rulemaking (Bray, 2014).

Although the case for AI-driven automation appears quite strong, AI tools are themselves far from perfect. They cannot be expected to address all deficiencies or errors in the public sector, nor will they be devoid of new sources of error of their own, especially if their forecasts are based on learning from data that already contain human biases. Moreover, when it comes to the application of AI in the public sector, a key question arises about their legality. It is one thing, after all, for AI tools to automate the selection and display of video choices on Netflix, but another for these tools to take the place of human judges and public administrators in making decisions that significantly affect people's lives. Prior to deciding whether to use algorithms to automate governmental decision-making, public administrators would do well to consider whether that use might pose any constitutional or administrative law roadblocks.

The Legality of the Automated State

An initial set of questions arises over whether government has the legal authority to substitute machines for humans. AI tools have prompted a series of legal concerns that largely stem from their black-box nature. Many of these legal concerns also arise, of course, with governmental decisions based on human judgment. Yet when concerns are targeted toward algorithms, particular worries have been expressed about the potential for machine learning to discriminate, obscure, and rob human officials of autonomy and discretion. In deciding whether to automate tasks using AI, then, public administrators will likely confront legal issues about accountability, procedural due process, transparency, privacy, and equality. We conclude that none of these five legal issues should act as any insuperable barrier to the responsible deployment of AI in the public sector, even deployment that might substitute automated decision-making for human judgment.

Accountability

The first issue centers on whether a governmental agency that replaces human-based processes with AI-driven automated systems would deprive individuals of a right to an accountable decision-maker (Busuioc, 2020). This concern is lessened when humans remain "in the loop" in AI systems—that is, when the systems only point humans to possible options and then the decision as to whether to take those options remains in human hands. A state supreme court, for example, sustained a trial judge's use of an algorithm rating tool in the criminal sentencing process because the scores generated by the algorithm were not the "determinative factor" in the sentence imposed by the judge (*State v. Loomis*, 2016).

Even human-out-of-the-loop systems, in which the outcomes of AI analysis fully replace human judgment, may still satisfy legal accountability principles. Under current constitutional law principles in the United States, for example, the nondelegation doctrine requires that a delegation of lawmaking powers must be constrained by an "intelligible principle" as to the basis for these systems' legal decisions. Is it possible that a delegation of decisional authority to an AI-based rulemaking system could run afoul of the nondelegation doctrine?

We think not, for two reasons. First, although the Supreme Court may well reinvigorate the nondelegation doctrine in the years ahead, the doctrine has in reality never placed a major constraint on administrative government and seems unlikely to do so in the future. The Court long ago determined, for example, that the legislative articulation of goals as vague as "public interest, convenience, and necessity" satisfy the intelligible principle test.

Second, and more importantly, any exercise of rulemaking authority by an automated system based on AI would be driven by what would of necessity be a highly intelligible principle, as machine-learning algorithms can only work if their goals are stated clearly and with mathematical precision. Thus, even under a potentially reinvigorated nondelegation doctrine in the years ahead, automation that relies on AI tools could not only pass muster under the intelligible principle test but will arguably make administration more accountable,

not less, given the precision with which machine-learning algorithm's objectives must be specified (Coglianese, 2021).

A related accountability issue arises when algorithms that are used by governments are designed and operated by private entities. In such instances, they could run afoul of the *private nondelegation doctrine*, which disfavors the outsourcing of governmental decision-making to nongovernmental entities. Although intuitively this doctrine might constrain governments in relying on contractors to develop AI systems, the principal rationale of the private nondelegation doctrine—avoidance of corruption—simply does not fit the context of machine learning. Private delegation has only been disfavored by the courts when private entities are likely to make decisions based on their own narrow self-interest instead of the broader public interest. Algorithms, however, are programmed to optimize objectives defined by the operators. As long as those operators act responsibly and are held accountable to the public through active governmental input and oversight, then the algorithms themselves pose no risk of corruption. AI-driven automation may even result in more accountable and faithful digital agents than human officials (Coglianese & Lehr, 2017).

Procedural due process

A second legal issue focuses on whether the principle of procedural due process requires a human decision, such that automation cannot lawfully take humans out of the loop (Citron, 2008). Procedural due process typically calls for government decision-makers who will listen, serve as neutral arbiters, and render reasoned judgments. Yet as much as these usual requirements might seem intuitively to call for a human decision, under prevailing constitutional principles what is demanded of procedural due process is supposed to be assessed using a balancing test comprising three factors: (1) the affected private interests; (2) the risk of decision-making error; and (3) the government's interests concerning fiscal and administrative burdens (*Mathews v. Eldridge*, 1976). AI-driven automated decisions would likely pass muster quite easily under this balancing test if an automated system would reduce errors and operate more efficiently. Indeed, due process might even eventually demand that government rely on algorithmic tools to achieve fairer and more consistent decisions than humans can deliver.

Moreover, despite the black-box nature of machine learning, due process expectations would likely be satisfied if automation is developed in a manner open to scrutiny (Coglianese & Lehr, 2019). Affected interests could be afforded the opportunity to interrogate design choices made by architects of these algorithms, and key design choices could be vetted through open processes with input from advisory committees, peer reviewers, and public comments or hearings.

Transparency

A third legal concern—transparency—stems from what has come to be considered a hallmark of best practices in public administration. One best practice—which has been called "fishbowl transparency" (Coglianese, 2009)—would imply that information about the design and operation of an algorithmic system be made publicly available and any key decision-making meetings about the algorithm should also be open to the public. Yet

fishbowl transparency is not absolute. Laws such as the Freedom of Information Act contain a variety of exceptions to required disclosures of public information, including for information involving trade secrets or related to law enforcement strategies. Some automated systems could legitimately fall under such exceptions.

An additional form of transparency—"reasoned transparency" (Coglianese, 2009)—demands that administrators provide reasons for their decisions. The use of machine-learning algorithms would seem problematic from the standpoint of reason-giving because their outputs cannot be intuitively understood or easily explained. Yet human decision-making can also be a black box, and human officials are not expected to furnish complete explanations of their decisions. Under prevailing reason-giving doctrine, administrators that rely on automation can likely satisfy their reason-giving obligations by explaining in general terms how their algorithms were designed to work and what data sources they are using (Coglianese & Lehr, 2019). Courts already defer to administrators' expertise in cases involving complex machinery or mathematical analyses, so it can be expected that courts will likely assume a similar deferential approach to evaluating an agency's reasons for decisions based on machine learning.

Furthermore, new technological advances continue to increase the interpretability of machine-learning models, opening up the algorithmic black box (Selbst & Barocas, 2018; Kearns & Roth, 2019). With time, administrators will have an even greater ability to make sense of the logic behind their decisions to all who are affected by algorithms.

Privacy

Because machine-learning algorithms thrive on large quantities of data, they raise a fourth category of legal concerns centered on privacy. Even when processed data includes sensitive or personal information, any privacy concerns should be as manageable with AI systems as they have been with large administrative data systems that do not rely on AI. Agencies already routinely handle an array of personal information, such as names, Social Security numbers, and biometrics. Protecting these data while concomitantly advancing agency goals is a task many agencies already have experience addressing.

Reasonably well-settled legal standards guide U.S. agencies in addressing privacy concerns. The Privacy Act of 1974 limits how agencies can collect, disclose, and maintain personal information. The E-Government Act of 2002 requires that agencies conduct privacy impact assessments when developing technology that implicates privacy concerns. Admittedly, growing calls exist for the federal government in the United States to adopt new data privacy laws, akin to the European Union's General Data Protection Regulation (GDPR) or to privacy laws adopted in some states, such as California. To the extent such new laws are adopted, they could necessitate greater attentiveness to privacy protections, but so far neither GDPR nor California's privacy law has created an insuperable barrier to the use of artificial intelligence.

Machine-learning algorithms do present a distinctive privacy concern in their ability to combine seemingly disparate, non-sensitive data to yield predictions about personal information, such as sexual orientation or political ideology. In the private sector, retailers have been able to use predictive analytics to infer private details from seemingly benign data, and then they have used those highly accurate inferences for targeted advertising. When it

comes to the public sector, the worry is that government officials could do the same, but for ill purposes. Still, such officials already are constrained from using this reverse-engineering potential of learning algorithms in ways adverse to individuals' interests because current law prohibits the "abuse of discretion" by government officials (Appel & Coglianese, 2021). As long as federal agencies use AI tools responsibly, there should seem to be no intrinsic privacy law impediment to that use.

Equal protection

Finally, automated systems raise legal concerns about equality. Biases in the existing data on which machine-learning algorithms train may simply become perpetuated, or perhaps even exacerbated, in AI-driven systems. Even when demographic details on race or gender are excluded from training datasets, algorithms might "find" a pattern based on these variables using the other variables in the dataset. Some research has documented the potential for racial and gender discrepancies in AI-based systems used for employment hiring (Bertrand & Mullainathan, 2004) and medical decision-making (Obermeyer et al., 2019). A widely circulated report published in ProPublica showed signs of racial bias in the (non-learning) algorithms in a system known as COMPAS, which is meant to provide an objective measure of recidivism (Angwin et al., 2016).

Bias obviously exists with human decision-making, and indeed human bias will often be an underlying source of bias in AI-driven systems when they are trained on data that accumulated in the past from human judgments (Mayson, 2019). But bias is also a concern with machine-learning algorithms, especially when the underlying training data are already biased. As a legal matter, when bias is intentional, it will clearly offend constitutional equality protections—whether in a system driven by humans or machines. But absent some independent showing of such intentional animus in the underlying objectives established for an AI system, it will be difficult for litigants to demonstrate unlawful discrimination by AI simply from its outputs.

One difficulty for individuals in protected classes who are adversely affected by an AI system will be to demonstrate that they were discriminated against *on the basis* of race (Coglianese & Lehr, 2017). This challenge will arise due to the inscrutability of machine-learning algorithms; they do not assign weights to or permit causal inferences about specific variables. Even when individuals within a protected class are treated worse than others, it is possible that the algorithm led to better outcomes for that class overall than would have been in a counterfactual situation. It might even be that any demographic variables in the dataset on which the algorithm trained were not actually influential at all in producing the outcomes. Individuals claiming algorithmic injustice will almost certainly need to show that the government's decision involved real reliance on a suspect classification. Such claims will face an uphill climb.

Another obstacle to sustaining an equal protection challenge to machine-learning systems will be the lack of "categorical treatment" in any adverse decisions produced through machine learning (Coglianese & Lehr, 2017). The U.S. Supreme Court has disapproved of administrative decisions on equal protection grounds when the government has afforded a categorical preference or disadvantage to certain classes. Such categorical treatment is

unlikely ever to arise with automated administration because an algorithm's objective function will presumably be defined in terms of some class-neutral outcome, even if underlying data contains human bias. As a result, we may expect few equal protection challenges to algorithmic administration to trigger heightened scrutiny. Furthermore, even if a court did apply heightened scrutiny to a machine-learning system, this might not lead it to find a violation of equal protection. After all, when administrators rely on machine-learning systems, they often do so to advance the kinds of compelling state interests which can withstand heightened scrutiny.

Summary

The five legal issues potentially implicated by governmental use of machine learning—accountability, procedural due process, transparency, privacy, and equal protection—do not appear to present any insurmountable barriers to the use of AI under prevailing law. These legal issues already arise with respect to governmental decisions based on human judgment or more conventional types of analytic tools. If the courts tolerate a status quo grounded on human biases and frailties, then they are unlikely to object categorically to automated systems that rely on digital algorithms, especially when these automated systems can be shown to overcome human limitations.

Obviously, the law may always change in ways that could lead to greater scrutiny of AI tools, as legal rulings and precedents may evolve alongside the increased use of algorithmic decision-making in government (Deeks, 2019). Indeed, some local jurisdictions have in recent years adopted restrictions on the use of AI-based facial recognition software by law enforcement personnel. Yet it is also quite possible that legal change in the future could work in favor of the adoption of AI tools. In some cases, we might anticipate that the responsible use of algorithmic tools will *enhance* principles of administrative justice and that eventually the use of these tools by government will come to be expected if not even legally required (Coglianese & Hefter, 2021). This would be especially so if data scientists continue to make progress in developing new algorithmic techniques that successfully address the various fairness, transparency, accountability, and other concerns that have been raised about the use of AI tools (Berk & Kuchibhotla, 2021; Kearns & Roth, 2019; Mullainathan 2019).

Nevertheless, in concluding that prevailing administrative law and constitutional law principles pose no intrinsic legal bars to the use of AI, we do not deny that some specific instances of automation will likely be challenged in court in the future. Already, algorithmic tools deployed by governments have found themselves subject to legal challenge (Coglianese & Ben Dor, 2021; Yoo & Lai, 2020). If administrators fail to exercise due diligence in their development of particular automated systems, they may well face court decisions disapproving of their use or finding that such use violates due process or is "arbitrary or capricious." As a result, when it comes to adopting AI in specific contexts for specific tasks, public administrators will need to take care whenever deciding whether to replace human decision-making with an automated system and, when they do, to take care in how they design and operate such systems. The key will be for public administrators to act responsibly when choosing automated administration.

Choosing Automated Administration

Decisions about the use of automation will need to be guided by a careful assessment of whether AI tools can lead to better processes and outcomes. Such evaluative judgments cannot be meaningfully made in the abstract but must be made within specific contexts and with respect to specific tasks. In some cases, machine learning will prove more beneficial than human decision-making, while at other times it will not. Despite the inherent context-specific nature of choices about automation, it is possible to outline relevant factors that ought to be considered. In general, the criteria for deciding whether to shift an existing human-driven process to one driven by artificial intelligence will fall into three categories: preconditions for the successful use of automation, the overall value of any such use, and considerations related to justifying automation so that it meets with acceptance by courts, legislative overseers, and the public.

Preconditions for use

To use machine-learning tools effectively, a public administrator will need to ensure the necessary preconditions for such use are met. Where the preconditions for artificial intelligence cannot be sufficiently met, automation should be avoided. The following four conditions are necessary even if not sufficient conditions for shifting from a human-driven process to automation.

1. *Adequate resources*. Administrators will need both the budgetary and human capital resources needed to build, test, maintain, and oversee an automated system. These systems require expertise, some of which could reside within a government agency but which might also need to be met through private contractors. Government agencies will also need adequate digital computing resources, such as data storage and digital processing capabilities. Both the stored data and computing processors will need adequate cybersecurity protections.
2. *Goal clarity and precision*. Machine-learning algorithms must be programmed to optimize a mathematically defined objective. Vague goals, such as fairness or reasonableness, cannot meet the mathematical precision needed. Fairness would need to be defined, such as by specifying that favorable outcomes generated by an automated system are to be distributed across racial and gender groups in ways that are proportionate to the distribution in an applicant pool or in society overall. In addition, any tradeoffs between values must be specified precisely. If the question is how much forecasting accuracy should be sacrificed for proportionality in outcomes, this tradeoff will need to be specified mathematically. In making such tradeoffs, agencies will need to have sufficient statutory direction and social input about how to define such tradeoffs.
3. *Data availability*. Machine learning works by identifying patterns within large quantities of data. If large quantities of relevant data are not available, then machine learning will not be an option for automating an agency task. Data may be lacking for

a variety of reasons. In some cases, they may exist, but only in paper form. Or they may be practically unavailable if disparate digital datasets are unable to be combined because they lack a common identifier to link data about specific businesses or individuals.

More conceptually, data may also be unavailable due to an insufficient number of repeated events around which a common set of variables could be collected. It is one thing to draw on data of individuals' DNA, for example, to determine whether a DNA sample from a crime scene matches that of a defendant. It is another to have a dataset that could determine whether a defendant was driving a red Corvette sports car through the intersection of Sixth and Main Streets in the evening of July 23rd (Rigano, 2019). It will obviously be difficult to find a large set of data for cases or tasks that are truly *sui generis*.

4. *External validity*. The data available for training a machine-learning algorithm must fit with and be applicable to the population that will be subjected to the outcomes of an automated system. If relevant factors change more quickly than an algorithm can be replenished with new data and retrained, then the algorithm can become "brittle"—in other words, lack external validity (Cummings, 2020). A machine-learning algorithm used for economic forecasting, for example, might work well under relatively stable economic circumstances but fail to generate accurate forecasts of employment during an unprecedented pandemic-induced recession. Of course, when circumstances become unusual, human judgment can also become brittle, so the need for external validity on the part of AI systems does not demand perfection but instead requires AI systems provide greater external validity than existing human-based alternatives.

The value of use

When the four preconditions above are met, AI-driven automation will be a plausible substitute for human judgment. None of these preconditions needs to be perfectly satisfied. But if they are not even minimally met in a given use case, it will not make sense to contemplate the use of automation.

In deciding whether to go forward in instances where the preconditions are met, public administrators should determine whether improvement over the status quo will be attained. Although the specific criteria for determining such improvement will depend on the specific use contemplated (Young et al., 2019), at least three general sets of impacts warrant consideration.

1. *Task performance*. Current human-based systems center on tasks, so one set of criteria should be guided by the objectives underlying those tasks: Would automation complete public administrators' tasks more accurately? Would it reduce the time it takes to complete these tasks? Would it be less costly, needing fewer employees or other resources? Would automation yield a greater degree of consistency in outcomes? Each of these questions can be asked in light of the current purpose of the human-driven system in use. The overarching test will be to determine whether automation would help public administrators better achieve their objectives.

2. *User or beneficiary impacts*. It will also be vital to attend to the effects of automation on the applicants, beneficiaries, or other individuals and businesses who use or are directly affected by a specific system. These impacts may already be part of a set of task objectives, but if not they should be carefully considered. How does the system treat those directly affected by its outputs? Do some portion of users suffer disproportionate adverse effects? Do they feel like the system has sufficiently served them well? In answering these questions, it is not necessary that an automated system be perfect—just better than the status quo. If the status quo of human-based systems necessitates that members of the public wait hours on the telephone to speak to someone who can assist them, a machine-learning chatbot might be much better in relative terms, even if it is not completely ideal. In this regard, it is notable that the online clearinghouse, eBay, deploys a fully automated dispute resolution software system that resolves disputes so satisfactorily that customers with disputes are reportedly more inclined to return to eBay than those who never had a dispute (Barton & Bibas, 2017).
3. *Societal impacts*. An administrator should also consider the broader societal effects of an automated system: How would automation indirectly affect others? Will remaining errors have broader consequences beyond the direct users? Errors might not have huge ramifications with, say, an automated customer service system used for answering commonly asked questions. But the ramifications to society overall could be profound for an AI system that automatically determines who receives commercial aircraft pilot licenses. Again, the key question to ask will be whether any broader societal effects would be better or worse than similar effects under the status quo.

In characterizing the value of an automated system across all three sets of impacts—task performance, direct impacts, and indirect impacts—it is important to be as specific as possible about the *degree of* performance improvements (or performance declines) resulting from a shift to automation. Automation will not necessarily fare equally well on each of these impacts. Decisions will need to be made about which impacts are more important. If automation proves to be less costly but slightly less accurate than the status quo, how important is accuracy for the use case at hand? How consequential are any errors that might remain with an automated system? It might be one thing for the U.S. Postal Service to tolerate some modest number of additional mistakes in letter sorting if automation lowers the costs of mail sorting dramatically. But it would be another matter altogether to make that same kind of tradeoff with respect to a system used for identifying safety risks on offshore oil rigs.

Justifying the use

Decisions to automate governmental tasks can pose risks of controversy and conflict. Agency officials should consider these risks and seek to manage them by building a justification for their use of automation. The justification will need to inform and satisfy legislative overseers, courts, and members of the public. The need for justification—and the risk of controversy—will likely be affected by two factors: (1) the degree to which automation overrides human decision-making, and (2) the stakes associated with an automated system's use.

Automation could provide just an *input* into human decision-making, as in the use at the heart of *State v. Loomis* (2016). Or automation could provide a *default* output that stands unless overridden by a human. Or an automated system could make the final *decision*

subject only to judicial review—essentially, a human-out-of-the-loop system. All things being equal, new uses of machine learning that only provide inputs into agency actions will be less likely to create controversy or be challenged in court than uses that create defaults or decisions (Coglianese & Lehr 2017).

In addition, the lower the stakes associated with the task performed by automation, the lower the risk of controversy or litigation. Among the lowest conceivable stakes will be purely internal staff uses within a public agency, such as when an agency's computer support staff uses a machine-learning algorithm as part of a chatbot that answers calls for password resets on office computers. That chatbot could even be designed to work with humans out of the loop, responding to requests entirely on its own, and yet controversy will be highly unlikely because the stakes are low (Mulholland, 2016). On the other hand, automated systems that are involved in the processing of applications for economically valuable licenses or permits by private businesses will have substantial stakes—and thus will pose risks of controversy or even litigation—even if they provide only inputs into decisions made by human officials.

When AI systems create defaults or decisions over high-stakes matters, public administrators can anticipate controversy to ensue or legal challenges to be filed. In these cases, they will need to take pains to demonstrate that they have engaged in careful planning and testing of these automated systems. Public notice and public engagement will often be appropriate in the development and design of automated systems in these circumstances, both to build confidence that may head off controversy and to reassure overseers that such systems have been designed and are being used responsibly.

Validation of automated systems will also be vital in justifying their use. Administrators should make validation data available to the public. They may also consider subjecting the design and performance of automated systems to external peer review or third-party auditing. Pilot programs could be established to run the algorithm in tandem with human decision-makers for a period of time to observe how it will operate in practice. Once the digital system has replaced a human-driven system, validation efforts should continue to occur early in its use before any irreversible loss of human capital. Future upgrades to the digital system will benefit from continued validation that each iteration improves on the one before—or at least does not present unacceptable side effects.

When contracting out for technical support and services in developing a machine-learning system, public administrators should scrutinize with care the sale pitches that contractors make (Dekel & Schurr, 2014), aware of the reality that their agencies, more than the contractors, will be held responsible for problems that arise with an automated system. In crafting contracts for AI services, administrators will do well to account for their need to access and disclose sufficient information about the algorithm, underlying data, and validation results to satisfy transparency norms (Coglianese & Lampmann, 2021). Overall, administrators should ensure that the procurement process will result in transparent outcomes for public administration, even if it involves proprietary software that would otherwise shield the algorithms from scrutiny.

Finally, public administrators should approach automation with humility, caution, and care, recognizing that AI systems are themselves creatures of human decision-making that can be prone to blind spots and biases. Because frailties of human judgment affect all human decisions—including decisions about whether and how to design and use automated systems—public administrators should remain vigilant. Such vigilance will also help in reassuring courts, legislative overseers, and the public.

Conclusion

A future of governance driven by AI-based automation naturally will give rise to concerns by judges, policymakers, and members of the public about how such use of digital tools will affect the efficacy, fairness, and transparency of governmental processes. But the status quo already relies on human "algorithms" that are far from perfect. If algorithmic automation can usher in a new public administration that—at least for specific uses—achieves better results at constant or fewer resources than current processes based on human decision-making, then both government administrators and the public would do well to support the use of artificial intelligence.

We conclude that the current legal system creates no insuperable obstacles to such public sector use of automation as a substitute for, or an augmentation of, current decision-making by human officials. Choices about whether and when to use AI tools will thus rest firmly on the shoulders of public administrators. These choices should be made with care, taking into due consideration the preconditions for the use of AI, the value of such use in terms of improving the status quo, and the need for public administrators to engage the public and ensure sufficient justification of their decisions about automation, especially when the stakes are high. Responsible, well-informed human decisions about the design and deployment of automation will shape whether AI can succeed in overcoming the limitations of human decision-making and improving governmental performance.

Note

* Although no universally accepted definition of AI exists, for the purposes of this chapter we focus on systems that rely on the use of machine-learning algorithms, both supervised and unsupervised. Such algorithms, which themselves come in a variety of forms, are increasingly making possible the automation of a range of previously human-centered activities, such as with the driving of automobiles and the reading of x-rays. These kinds of algorithmic tools are sometimes described with a variety of phrases, including "big data," "predictive analytics," "deep learning," "reinforcement learning," "neural networks," "random forests," and "natural language processing." In the context of public administration, the use of such tools would constitute one manifestation of what has been described variously as smart or data-smart government (Goldsmith & Crawford, 2014; Noveck, 2015) or moneyballing government (Nussle & Orszag, 2015). For discussion of the relevant characteristics of the type of machine-learning algorithms we have in mind, see Coglianese & Lehr (2019), Anastasopoulos & Whitford (2018), and Coglianese & Lehr (2017).

References

Agrawal, A., Goldfarb, A., & Gans, J. (2018). *Prediction machines: The simple economics of artificial intelligence*. Boston, MA: Harvard Business Review Press.

Alhola, P., et al. (2007). Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric Disease & Treatment 3*, 553.

Anastasopoulos, L. J. & Whitford, A. B. (2018). Machine learning for public administration research, with application to organizational reputation. *Public Administration Research & Theory 29*(3), 491–510.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Appel, S., & Coglianese, C. (2021). Algorithmic governance and administrative law. In Woodrow Barfield (Ed.), *Cambridge handbook on the law of algorithms: Human rights, intellectual property, government regulation*, 162–181. Cambridge, England: Cambridge University Press.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy 58*, 328–346.

Baekgaard, M., et al. (2019). The role of evidence in politics: Motivated reasoning and persuasion among politicians. *British Journal of Political Science 49*, 1117.

Barton, B. H., & Bibas, S. (2017). *Rebooting justice: More technology, fewer lawyers, and the future of law*. New York, NY: Encounter Books.

Berk, R., & Kuchibhotla, A. K. (2021). Improving fairness in criminal justice algorithmic risk assessments using conformal prediction sets. arXiv preprint. https://arxiv.org/abs/2008.11664.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review 94*, 991.

Bray, D. A. (2014). An update on the volume of open internet comments submitted to the FCC. *Federal Communications Commission*. https://www.fcc.gov/news-events/blog/2014/09/17/update-volume-open-internet-comments-submitted-fcc.

Bureau Labor Statistics (BLS). (2019). Automated coding of injury and illness data. https://www.bls.gov/iif/autocoding.htm.

Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review 81*(5), 825–836.

Chenok, D., & Yusti, C. (2018). The future has begun: Using artificial intelligence to transform government. *IBM Center for The Business of Government*. https://ourpublicservice.org/wp-content/uploads/2018/01/0c1b8914d59b94dc0a5115b739376c90-1515436519.pdf.

Citron, D. K. (2008). Technological due process. *Washington University Law Review 85*, 1249.

Coglianese, C. (2009). The transparency president? The Obama administration and open government. *Government 22*, 529.

Coglianese, C. (2021). Administrative law in the automated state. *Dædalus 150*, 104–120.

Coglianese, C., & Ben Dor, L. M. (2021). AI in adjudication and administration. *Brooklyn Law Review 86*(3), 791–838.

Coglianese, C., & Hefter, K. (2021). From negative to positive algorithm rights. *William & Mary Bill of Rights Journal* forthcoming.

Coglianese, C., & Lampmann, E. (2021). Contracting for algorithmic accountability, *Administrative Law Review Accord 6*, 175.

Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal 105*, 1147.

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review 71*, 1.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences 24*, 87.

Cummings, M. L. (2020). The surprising brittleness of AI. *Women Corporate Directors*. https://www.womencorporatedirectors.org/WCD/News/JAN-Feb2020/Reality%20Light.pdf.

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review 119*, 1829–1850.

Dekel, O., & Schurr, A. (2014). Cognitive biases in government procurement—An experimental study. *Review of Law & Economics 10*, 169.

Eberhardt, J. L. (2019). *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. New York, NY: Viking.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by algorithm: Artificial intelligence in federal administrative agencies*. Administrative Conference of the United States. https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology 1*, 288.

Food & Drug Administration (FDA). (2011). FDA Commissioner's fellowship program. https://www.fda.gov/media/83569/download.

Fisher, M. (2020). Coronavirus "hits all the hot buttons" for how we misjudge risk. *New York Times*. https://www.nytimes.com/2020/02/13/world/asia/coronavirus-risk-interpreter.html.

Forrest, J. (2005). The Space Shuttle Challenger disaster: A failure in decision support system and human factors management. *Decision Support Systems Resources*. http://dssresources.com/cases/spaceshuttlechallenger/index.html.

Freund, C., & Özden, C. (2008). Trade policy and loss aversion. *American Economic Review* 98, 1675.

Gates, S. W., et al. (2010). Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate 13*, 369.

Goldsmith, S., & Crawford, S. P. (2014). *The responsive city: Engaging communities through data-smart governance*. San Francisco, CA: Jossey-Bass.

Halpern, S. D., Truog, R. D., & Miller, F. G. (2020). Cognitive bias and public health policy during the COVID-19 pandemic. *Journal of the American Medical Association 324*, 337–338.

Hino, M., Benami, E., & Brooks, N. (2018). Machine learning for environmental monitoring. *Nature Sustainability 1*, 583–588.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin 121*, 43–64.

Janis, I. (1972). *Victims of groupthink*. Boston, MA: Houghton, Mifflin.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. New York, NY: Oxford University Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*, 237–293.

Kogan, N., & Wallach, M. A. (1967). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology 3*, 75–84.

Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science 57*, 1827.

Lai, A. (2021). Artificial intelligence, LLC: Corporate personhood as tort reform. *Michigan State Law Review 2021*, 599.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality & Social Psychology 37*, 822.

Lewis, M. (2021). *The premonition*. New York, NY: W.W. Norton & Company.
Light, P. C. (2014). A cascade of failures: Why government fails, and how to stop it. *Center for Effective Public Management at Brookings*. The Brookings Institution. https://www.brookings.edu/research/a-cascade-of-failures-why-government-fails-and-how-to-stop-it/.
Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality & Social Psychology 37*, 2098.
Mathews v. Eldridge, 424 U.S. 319 (1976).
Mayson, S. (2019). Bias in, bias out. *Yale Law Journal 128*(8), 2218.
McKinsey Global Institute. (2018). Applying AI for social good. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/applying%20artificial%20intelligence%20for%20social%20good/mgi-applying-ai-for-social-good-discussion-paper-dec-2018.ashx.
Mullainathan, S. (2019). Biased algorithms are easier to fix than biased people. *New York Times*, December 6. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html.
Mulholland, J. (2016). Chatbots debut in North Carolina, allow IT personnel to focus on strategic tasks. *Government Technology*, October 11. https://www.govtech.com/computing/chatbots-debut-in-north-carolina-allow-it-personnel-to-focus-on-strategic-tasks.html.
National Institute of Standards and Technology (NIST). (2021). Artificial intelligence. https://csrc.nist.gov/Topics/technologies/artificial-intelligence.
Neustadt, R. E., & May, E. R. (1986). *Thinking in time: The uses of history for decision-makers*. New York, NY: Free Press.
Noveck, B. S. (2015). *Smart citizens, smarter state: The technologies of expertise and the future of governing*. Cambridge, MA: Harvard University Press.
Nussle, J., & Orszag, P. (2015). *Moneyball for Government*. New York, NY: Disruption Books.
Nutt, P. C. (1999). Surprising but true: Half the decisions in organizations fail. *The Academy of Management Executive (1993–2005) 13*, 75–90.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*, 447.
Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
Panigrahi, B. K., et al. (2018). Nature inspired computing. *Advances in Intelligent Systems & Computing 652*, 203.
Pascalis, O., et al. (2005). Plasticity of face processing in infancy. *Proceedings of the National Academy of Sciences 102*, 5297.
Peters, R. (2006). Aging and the brain. *Postgraduate Medical Journal 82*, 84.
Pretorius, A., & Parry, D. (2016). Human decision making and artificial intelligence: A comparison in the domain of sports prediction. *Annual Conference of the South African Institute of Computer Scientists and Information Technologists 32*, 1–10.
Rehavi, M. M., & Starr, S. B. (2014). Racial disparity in federal criminal sentences. *Journal of Political Economy 122*, 1320.
Report of the Presidential Commission on the Space Shuttle Challenger Accident. (1986). https://history.nasa.gov/rogersrep/v2appg.htm.
Rigano, C. (2019). Using artificial intelligence to address criminal justice needs. *NIJ Journal*, January. https://www.ojp.gov/pdffiles1/nij/252038.pdf.
Schuck, P. H. (2014). *Why government fails so often: And how it can do better*. Princeton, NJ: Princeton University Press.

Selbst, A. D. & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review 87*, 1085.

Slavitt, A. (2021). *Preventable: The Inside Story of How Leadership Failures, Politics, and Selfishness Doomed the U.S. Coronavirus Response*. New York: St. Martin's Publishing Group.

Shrestha, Y., et al. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 6*, 66.

State v. Loomis, 371 Wis.2d 235 (2016).

Sunstein, C., Kahneman, D., & Sibony, O. (2021). *Noise*. New York, NY: Little, Brown Spark.

TRAC. (2011). Social security awards depend more on judge than facts: Disparities within SSA disability hearing offices grow. https://trac.syr.edu/tracreports/ssa/254/.

Tschandl, P. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology 20*, 938.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty. *Science* 185, 1124.

Vaughan, D. (2016). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. Chicago, IL: University of Chicago Press.

Wyden, P. (1979). *The Bay of Pigs: The untold story*. New York, NY: Simon and Schuster.

Yoo, C. S., & Lai, A. (2020). Regulation of algorithmic decision-making in the United States. *Journal of Law & Economic Regulation 13*, 7.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management & Governance 2019*, 301–313.

CHAPTER 24

TRANSPARENCY'S ROLE IN AI GOVERNANCE

ALEX INGRAMS AND BRAM KLIEVINK

Transparency in Governance

As public decision makers continue to address the governance challenges of artificial intelligence, they will almost certainly have to address the question of how to integrate architectures of transparency into the design of artificial intelligence systems. Transparency is closely related to the concepts of information, communication, and knowledge and is therefore an integral part of the design of governance systems. Governance systems depend on coordination and communication of information by decision makers to whom such information must be to some degree transparent by virtue of the need for informed knowledge, planning, and action (in practice, governance systems give some decision makers more transparency than others). Governance systems also depend on transparency for public information. Such systems are oriented around producing valuable outputs for society in the form of information and knowledge on finance, service contracting, or citizen satisfaction that can subsequently be used to improve services such as education, transportation, or health care.

In these basic ways, governance *of* transparency and governance *by* transparency are vital for society. In this chapter we focus on governance *of* transparency, which is about what decision makers do to support transparency in the use of AI for performing governmental functions. Governments have always tried to protect state secrets, publicized information about their programs, and developed legal systems such as Freedom of Information laws that stipulate how citizens can request government information that is not yet in the public domain. However, the challenge of governance in transparency design runs deeper than this. Transparency is not just production and regulation of information. It is, essentially, a multi-actor relational process of making information visible and comprehensible where information is produced, shared, published, and protected by different organizations or actors. As we will discuss later in this chapter, this has important implications for the role of transparency in governance of AI. Transparency also connotes a kind of quality to information that is about the visibility of information but also its *meaningfulness* or *actionability* to

involved actors. In other words, transparency is information that conveys something about a real state of human affairs with relevance to future situations and decisions. It should be quite clear from this definition of transparency that—despite its value to public communication and services—it is a difficult thing to manage and integrate into the design of governance systems for artificial intelligence. This is because AI systems can be overwhelmingly complex to a degree that their interconnectedness with human decision making often becomes tenuous and fuzzy. Dealing with this challenge is important for the way government and public servants use AI in public policy and public services and how governments determine what transparency demands of them, which transparency goals are to be met, toward whom, and how.

To address this challenge in this chapter, we investigate the scholarly debate on transparency in governance and, in particular, we explain why transparency is a difficult thing to manage. An important second component of the chapter is to then combine these debates about transparency with emerging debates on the governance of artificial intelligence. In joining these things together, it should become clear that the application of artificial intelligence in the public sector presents engineers and systems designers with a new set of difficult transparency problems. We will survey the state of the art in both these fields and offer an overview of existing approaches to transparency, synthesizing the various views that exist in the literature. In so doing, we aim to promulgate a richer transparency debate and practice, beyond a simple "transparent-or-not" approach to AI. We conclude with arguments and an example of smart energy meters for incorporating these difficult conceptions in an integral approach when designing transparency in and for AI systems. The use of artificial intelligence in the public sector presents a complex governance task and we argue that only a versatile approach, responding to different institutional and actor contexts, will deliver the kinds of context-dependent transparency solutions needed for AI systems.

Transparency crops up frequently in theories about public governance. It is normally considered a positive force for encouraging *good* governance. But transparency is also a multifaceted concept. This is evident, firstly, in its close association with the facts–values dichotomy: it is related both to scientific concepts (such as evidence, perception, and knowledge) and to moral concepts (such as openness, honesty, and corruption). The multifaceted nature of transparency led to Hood and Heald (2006) describing transparency as a "magic concept" that can mean a lot of different things to different people. It is a concept that is widely manipulated by political actors or seen as a panacea for many of government's governance problems. In attempts to navigate this ambiguity, scholars of transparency have therefore created typologies of different kinds of transparency, describing how they function and the kinds of relationships, actors, and information sharing processes that they entail.

Nearly two centuries ago, early theorists of liberal democracy such as John Stuart Mill (1806–1873) and Alexis de Tocqueville (1805–1859) recognized that transparency underpins liberal democracy and argued that government relies on citizens getting access to information and using it to discuss public issues that affect their lives and wellbeing (Ingrams et al., 2020). While in this sense transparency is important for public deliberation and understanding, its role extends far beyond democratic governance to encompass smooth functioning and statutory requirements at all levels of governmental processes and decision making. According to Heald (2012), transparency operates between various internal levels of governance too, such as in vertical relationships between managers and staff or political appointees and civil servants. Heald contrasts this kind of vertical transparency with the

horizontal forms of transparency that constitute democracy and give citizens information about the internal actions of government as well as giving government information about the citizens themselves (such as service use rates, satisfaction levels, or demographics) and other third-party actors.

Transparency of AI

Transparency of AI shares essential features of traditional transparency in governance; the goals of transparency here are essentially the same in terms of giving vertical and horizontal insights to internal decision makers and citizens that can improve governance. However, transparency of AI also expresses some new puzzles about transparency because of the way AI processes work and the kinds of interactions they introduce with human decision makers. The challenge of designing transparency of AI is inherently more elaborate and complicated than an analogue form of transparency, such as a public procurement records repository or text minutes taken from parliamentary proceedings. This is due to the copious amounts of information used as input in training and/or applying AI models and technologies, due to the complexity of the AI itself, in terms of software, code and the incomprehensible nature of the inner workings of algorithms. Finally, it also stems from the increasingly complex ways AI augments human processes and is used in institutional settings.

In its simplest form, transparency of AI is often understood as the provision of public information about what an organization (e.g., departments, municipalities, inspection agencies) does toward stakeholders that are somehow affected by the use of algorithms, or have the right to know (e.g., democratically elected bodies, or the general public). Initiatives on Explainable AI (XAI) or the enactment of policies that enforce open source, auditable code, help here because they technically lead to more information about how the AI works. Yet, this offers a form of *fishbowl* transparency rather than *reasoned* transparency, meaning that the latter concerns the "why?" or the motivations lying behind programming decisions rather than revealing only the external facts or actions without context as in the former (Coglianese & Lehr, 2019). Transparency is not just a feature of the technology that speaks for itself without human context. As suggested by our definition of transparency above, transparency depends on whether information reveals the intentions of the designers and whether the information is actionable for the recipient of the information, and the user.

In situations where transparency on the actual workings of an AI-based system to the general public is required, this has a significant impact on what kinds of information need to be provided and what kinds of institutional and organizational capacities help to make the information actionable. It might be required that experts (either non-technical users or technology professionals) within the organization can understand what AI-driven intelligence they work with. Or it might only be necessary that experts outside of the organization such as auditors are able to access and understand what AI is used and how. Each of these audiences presents a different way that reasoned transparency should be achieved taking into consideration the background information of motivational context needed to understand how the AI is being used and to what ends.

AI and human decision-making

Most of our theories about transparency in governance are premised on the idea that information produced by human actions can reveal key aspects of the decision making behind those actions. In this sense, AI can be superior in transparency to human decision making when there is a digital trail of computer programming steps involving large amounts of data (Ku & Leroy, 2014; Skaug Sætra, 2020). The inner workings of the human mind are impossible to render transparently this way, even if it were desirable to do so. Thus, for example, machine learning approaches can efficiently annotate large corpuses of text and provide more fine-grained information than human approaches (Haunss et al., 2020).

But AI, while analogous to forms of human cognition, processes information in different ways that undermine such AI–human equivalence. According to König and Wenzelburger (2020), this means that while AI can reduce information sharing problems for citizens and government in some respects, it can increase them in others, particularly when it comes to the problem of rendering meaningful and actionable information about AI decision making. Firstly, there are specific types of legal challenges resulting from the differences between AI and human management of information. For example, in the words of Gellert (2022, para. 1), "data protection law is grounded in the logic of knowledge communication, which stands in stark contrast with machine learning, which is predicated upon the logic of knowledge production, and hence, upon different definitions of data and information."

Secondly, transparency of basic computer programming steps and decision outputs may be possible, but computers cannot explain or justify their behavior in the ways that humans can or give account of their reasons. AI can change an uncertain evidence base into a more certain one using statistical prediction, but this type of certainty may obscure other critical ingredients of an AI decision making process in the way it handled the original data, perhaps by ignoring or weighting some aspects over others (Hartmann and Wenzelburger 2021) or promoting correlation over causation. The invisibility and autonomousness of algorithmic calculations presents an accountability problem that distinguishes it from human accountability (Dauvergne, 2020; Janssen & Kuk, 2016). While a human-led governance system can set some parameters or rules, such decision making is itself influenced by politics and human error and could lock in an algorithmic structure in a way that a purely human processes subject to changing human desires or situations would not (Cihon et al., 2020; Firlej & Taeihagh, 2021). Disclosure devices for transparency that are built into such hybrid systems are neither entirely human nor entirely computational, which has important implications because different degrees of computer–human interaction create different transparency barriers and influence the kind of transparency that is possible.

Transparency challenges of AI

While the governance challenges of transparency are novel due to the way AI–human decision-making systems work, public administration scholars have for several decades already been studying a similar set of problems regarding the relationship between digitalization (the growth of digital technologies and their displacement of analogue technologies) and transparency. Among major properties of digitalization that—similarly to artificial

intelligence—have consequences for governance of transparency are the fact that digitalization increases the amount and complexity of data available to government, that the internet provides for rapid and geographically widespread data sharing, and the way that cloud computing offers immense capacity for storability and versatility of data.

As a result of technology changes—which are often labeled collectively as "Big Data"—choices for public administrators in how they collect and present information have exponentially increased (Van der Voort et al., 2019). These properties of digitalization increase and alter the kinds of things that individuals and organizations can do, a phenomenon which is sometimes referred to as the "affordances of technology" (Zammuto et al., 2007). Affordances, as the name suggests, afford new economic, social, and political opportunities to technology users. The rise of artificial intelligence and its incorporation within existing transparency regimes has increased the potential for such affordances. This can have some genuine benefits for transparency. Automated processes, which are an integral part of artificial intelligence, can increase transparency because the computer code that programs the automated decision-making processes is traceable, replicable, and may be part of an automated learning loop that can be observed. But, simultaneously, the same properties of AI that support useful affordances—the complexity of programming and the consequences of decision making by artificial intelligence—introduce new challenges (Gillis & Spiess, 2019). City dashboards, for example, which offer affordances for gathering city data and projecting trends and graphics in real time in areas such as crime, crowdedness of shopping malls or transportation, and public health risks, "need to be complemented by mechanisms supporting citizens' engagement, data interpretation, governance and institutional arrangements" (Matheus et al., 2020, para. 1). AI has transparency implications for the managers and staff who are employing AI tools to improve their work processes. They employ the tools because of their affordance but, by the same token, the ability of AI to act independently can make it technically challenging, and the matter of management becomes a question of balancing the new opportunities with the concomitant costs.

Such costs affect all levels of the organization from planning and budgeting to media relations and communication, but they also can come in the form of barriers to implementing transparency. Burrell (2016) suggested that there are three specific types of transparency barriers:

1. *Intentional*: Barriers that involve the deliberate covering up or obfuscation of information to avoid personal or organizational costs associated with revealing the information.
2. *Illiterate*: Barriers stemming from a lack of understanding or technical knowledge on the part of public audiences of transparency. In this case the fault lies with the ability of audiences to understand information rather than actions taken by the providers of information.
3. *Intrinsic*: Barriers resulting from mismatches in how technological and human systems process information such as when information systems render information that is lacking is sufficient detail or information is prioritized in ways that are not as useful for users of technology.

Further, Burrell argues that the problematic degree of any of these three barriers derives from three kinds of public sector phenomena: secrecy, outsourcing, and complexity. To

Table 24.1 Barriers to transparency in governance of AI

	Intentional	Illiterate	Intrinsic
Secrecy	Energy companies protect smart meter intellectual property or government surveillance and security interventions.	Obfuscation of the AI algorithms may occur as energy companies or regulators are incentivized to show performance in the best light possible.	If critical algorithmic information is hidden, learning processes for AI–human understanding will be harmed.
Outsourcing	Regulatory goals compete with commercial incentives of energy providers.	Incentives of energy providers do not promote dedication to interpretability of their algorithm for regulators or consumers.	Information asymmetries resulting from lack of experience from energy companies leads to poor decisions about transparency for users.
Complexity	Energy performance fluctuates from place to place and smart algorithms are computationally dense, which facilitates cheating.	Smart algorithms are computationally dense, which makes it difficult for homeowners to use them to their benefit.	Complex algorithms may not self-correct for the equally complex behavior of households and energy use.

illustrate further, Table 24.1 shows what barriers could hypothetically occur in the case of household energy smart meters.[1] In this example, the AI is of a relatively simple kind with a limited number of synchronous tasks, and which is also focused on a narrow and stable set of data. The smart meters measure how much energy is being used by a homeowner and shares the information with energy providers. By making the information transparent to energy users and suppliers, policymakers estimate that the meter could help reduce energy waste. The meters also can switch off energy inflows if energy bills are not paid on time.

Transparency could be limited by secrecy, either of deliberate kinds (e.g., blocking access through intellectual property laws) or hiding the information. On the government side, surveillance could lead to secretive activities of government aimed at national security initiatives, such as collecting names or addresses of energy users. Secrecy can also result when energy companies or government regulators use their own discretion to partially limit or alter information to present it in a more positive light. In the last column of the table, intrinsic kinds of barriers result from the way the technology itself works. In this case, secrecy could occur not through human intentionality per se but through the black box character of the AI. Algorithmic models used to run AI can require specialized knowledge and include many inscrutable and nonintuitive features that enable secrecy and create accountability gaps when inscrutable information about an AI decision-making process is separated from its original design. When it comes to outsourcing, the challenge here for transparency is enabling information sharing when organizational boundaries between government and contracting organizations limit data sharing, even if the data sharing is in the public interest. Contactors may intentionally withhold information if they believe that publishing the information harms their market position. Third-party contractors are increasingly used for AI tools, and, purely from a technical perspective necessary for reasoned transparency,

this makes understanding and explaining increasingly difficult for administrators or citizens who are not (perhaps only later once the software has been implemented) involved in the application and use of the AI. In terms of intrinsic barriers, long-term dependency on contractors allows them to benefit from the information asymmetry that puts them in a position of power in a crucial area of society, such as energy consumption. Information asymmetries are also evident in the last row of the table: complexity enables intentional obfuscation of data for financial gain or it simply becomes too complex for consumers to understand it in such a way that they can manage their energy use efficiently. In such ways, AI is subject to complex mediation processes between software, engineers, and users that present barriers for turning a complex array of information into intuitive information that forms the basis for transparency.

AI and public values

The matter of AI and the new challenges it presents for transparency are often described as a challenge for *public values*. AI has consequences for how well public organizations perform on public values that they are expected to deliver in democratic societies, such as transparency, as well as related values like accountability and trust (Skaug Sætra, 2020). Public values are values that, "embody the prerogatives, normative standards, social supports, rights and procedural guarantees that a given society aspires to provide to all citizens" (Bozeman & Sarewitz, 2005, p. 122).

Technologies should aim at such public value creation in addition to making government run more efficiently (Panagiotopoulos et al., 2019). But AI tends to raise new questions about risk and tradeoffs to public values and brings these tensions into higher relief (Andrews, 2019; Henman, 2020; Panagiotopoulos et al., 2019). Further, the need for cross-sector collaborations to manage AI adoption in the public sector threatens public values (Andrews, 2019). This occurs because technology companies are often driven by different values from the public sector and are not subject to the same accountability requirements as governments, meaning that AI applications may remain obscured from public vision. The quest for greater efficiency and cost saving by using the latest systems available in the market may crowd out the call for transparency in cases where accuracy of AI decision-making depends on analytics that are highly effective at what they do but are simultaneously too complex to be rendered in an explainable, or even interpretable, way. Another risk for AI, according to Vogl et al. (2020), is that value tradeoffs become more explicit when clashes between values, such as privacy, are prioritized over transparency. Clearly, both values—privacy and transparency—are necessary. Data sharing can be too much and of the wrong type of personal information, which presents a risk for privacy. Similarly, according to Cofone (2019), secrecy in some circumstances (e.g., military intelligence, sensitive areas of parliamentary deliberation) or demographic factors in algorithms (e.g., economic or health-related backgrounds) is necessary to counterbalance absolute transparency in all areas of government.

In sum, AI systems can threaten public values, such as transparency, and they also present complicated problems about necessary tradeoffs between transparency and other important public values, such as privacy and efficiency. Harrison and Luna-Reyes (2022, p. 11) say that without addressing the public values challenges of AI, "the problems with issues of

algorithmic opacity will continue to jeopardize government's traditional values for transparency, explainability, and accountability."

Approaches to Governance of AI Transparency

Given the transparency challenges of AI and their threat to public values, many transparency scholars have argued that AI adoption in the public sector calls for new governance mechanisms and a need for the protection of public values to be institutionalized (Agarwal, 2018; Gorwa et al., 2020; König & Wenzelburger, 2020). Indeed, a variety of changes to public sector governance are already well underway as a result of AI adoption (Vogl et al., 2020). Without a careful and broad public discussion about these possibilities, our current governance structures will not keep pace with the exponentially rising growth of AI and will quickly become redundant or unsuitable to the new challenges.

The current debate is characterized by four main types of approaches: (1) constructivist, (2) democratization, (3) legal reform, and (4) capacity building. While these four approaches do not comprehensively cover all theories in the literature, they do capture the major differences and debates. As we will describe in more detail below, each approach has a distinct idea of the most critical values and different normative conceptualizations of AI transparency challenges and the technical approaches as to how the challenges should be tackled. By comparing these differing approaches, we highlight their key insights and show how each contributes to different but critical perspectives on transparency of AI that inform decision making on how to realize transparency in governance of AI.

Constructivist approaches

Constructivist approaches emphasize the fact that transparency is a somewhat misunderstood quantity, and that more needs to be done to understand the construction of transparency systems and their limitations, constrained by power and subject to political behavior. According to Latour (1987), black box algorithms are essentially self-contained amalgamations of human relationships created through an algorithm's programming process. If this conceptualization is the starting point, the quest to completely open up such black boxes must be misconceived because black boxes are accessed through the human relationships that led to them in the first place rather than through the artefacts (scripts, computer code, rules, etc.). Ultimately then, it is in those relationships and their replication in new networks of humans and technologies that the effort at seeing, understanding, and, ultimately, governing can be found.

Ananny and Crawford (2018) argue that it would be virtually impossible to reconstruct all the human decision-making points that would render AI systems completely transparent. Instead, it is better to look *across* systems, rather than inside them, because human

and technological systems actually *enact* complexity rather than contain it per se. Ananny and Crawford (2018, p. 984) argue that, "if we recognize that transparency alone cannot create accountable systems and engaging with the reasons behind this limitation, we may be able to use the limits of transparency as conceptual tools for understanding how algorithmic assemblages might be held accountable."

For constructivists, transparency will not by itself correct governance challenges of algorithms. Rather from a constructivist view of transparency, the important question is: "what kind of claims can be made to *understand* an actor-network, and how is this understanding related to but distinct from simply *seeing* an actor-network?" (Ananny and Crawford, 2018 p. 983, our emphasis). Constructivists "argue that if the 'culture of accountability' is to adapt to the challenges posed by new and emerging technologies, the focus cannot only be technology led. It must be focused on understanding the 'why?' of decisions and be interrogative of the governance choices that are made within organizations, particularly those vested with public functions at the international and national level" (McGregor, 2018, p. 1079).

Democratization approaches

According to König and Wenzelburger (2022, p. 7), "All in all, decisions count as 'good' according to the procedural standard of democratic legitimacy if they are the result of commonly accepted rules and procedures that regulate the use of public power." Democratization approaches suggest that AI represents a shift in the collective way that decision-making is made, and the risks entailed. Opening up algorithms to greater input from stakeholders (the "public") is therefore the best way to achieve transparency. By showing the work of AI decision making programming and inviting feedback, policymakers can, in principle show that they value democracy (Ingrams, 2020; Strandburg, 2019).

The solution to transparency of artificial intelligence in the democratization approach is an indirect one that focuses not on the design of algorithms per se but on the social environment—civil society organizations mainly—of governance of artificial intelligence, which constitutes a pathway for algorithmic transparency (Criado et al., 2020; Ingrams, 2019; Sun & Medaglia, 2019). While there is an especially important role for civil society intermediaries to offer ethical and expert input on the artificial intelligence's development (Williamson, 2014), transparency about who is involved in the design of artificial intelligence reveals that there is a diversity of stakeholders from individual entrepreneurs, users of technologies, and businesses (McGregor, 2018; Nicholson Price II, 2017).)

Capacity-building approaches

Capacity-building approaches focus on the lack of skills, expertise, and knowledge in public organizations (Klievink et al., 2017). Once these capacity and skills shortages are solved, the other challenges, such as transparency, will begin to be more manageable. More fundamental reform is probably not required. For some scholars, such as Harrison and Luna-Reyes (2022), capacity is a necessary starting block for AI transparency. They

say that "substantial human oversight and domain expertise would seem to be minimal requirements for AI system development. The goal should be an accumulated track record of successful AI development in low-risk applications that provide test beds for experimentation" (p. 20). The need for technical capacity and skills is also recognized in organizational checklists for effective governance of AI. For example, Wirtz and Müller (2019) provide the following checklist:

1. Codification of ethical AI standards and regulations and monitoring their enforcement.
2. Setting an AI agenda that defines targets, field of application, and a roadmap for employment.
3. Establishing limits and boundaries for AI usage and avoiding autonomous decision making.
4. Enhancing computer knowledge and AI-specific skills within the organization.
5. Providing insights to data acquisition and processing and creating verifiable AI algorithms.
6. Detecting options to automate administrative routine processes by means of AI.
7. Enlarging the working capabilities of staff by AI usage.

Legal approaches

Legal approaches focus on the role of new laws or the reapplication and reinterpretation of existing laws as the path toward making AI transparent. Principles of transparency are central to administrative law, and it is the unique qualities of AI that seem to challenge this because they lack "an interpretive ability to describe this optimization in conventional, intuitive terms" (Coglianese & Lehr, 2017, p. 1207). In contrast, the jurisprudence concept of *reason giving* requires decision makers to be thorough and probing, and to disclose basic rules. To support this, AI tools, such as visualizations, can accompany outputs to show how they work, thereby at least giving intuitive explanations (Coglianese & Lehr, 2017). Reasons for an algorithm's operations should be given at both *individual* and *group* levels so that effects of specific individual cases, as well as potentially on other affected group members, can be addressed (Coglianese & Lehr, 2017). Another key dimension of reason giving for Coglianese and Lehr (2017) is the *importance* and *functional form* of input variables. For example, some programming decisions give strong importance to some demographic variables (such as age or gender) and may assign a different function to those categories such as weighing them more positively or negatively.

An important concept for the legal approach is *explainability* (Strandburg, 2019). For users of delegated and distributed systems explainability is especially important because this is where simple automated systems can have an advantage over human decision-making. The former allows users to document an entire decision-making logic. According to Strandburg (2019, pp. 1872–1873), requirement for explainability includes five elements: (1) separation of decision criteria into automated and non-automated aspects, (2) definitions of the decision criteria, (3) definitions of outcome variables used for decision criteria, (4) definitions of feature variables to be used as factual evidence in automated decision criteria assessments, and (5) combination schemes governing how adjudicators combine automated assessments with other relevant information.

A proposed integrated governance approach to transparency of AI

Transparency is a multifaceted concept tested to the extreme by the new challenges of AI. Each of the four approaches to transparency of AI gives a different diagnosis of what the solution could be for governance of AI. They also have slightly different takes on what the main problem is in the first place. For constructivists, the problem is the mistaken notion that information can be a politically neutral tool for guiding political decisions. In contrast, the democratization approach sees the political exclusivity of decision making about AI systems as the root of the problem. By the same token, the solutions are also differently prescribed. To compare the same two approaches, constructivists prescribe shifting attention to the individuals and their relationships to understand what motivates and shapes the function of algorithms used to construct AI systems. In contrast, supporters of a democratization approach prescribe more involvement of citizens in the process of developing AI systems.

At a challenging time for the governance of AI, different ideas about the best path forward for transparency are inevitable. But there is a need for an integral approach, incorporating legal, organizational, and technical elements, and requiring the right capabilities to be acquired (if not developed). Such an approach is not without limitations. Like any solution, it is challenged by problems of control ("classic" policy implementation and management problems), as well as overly strong institutionalization that might hamper innovation and thereby the learning that is much needed at this stage of rapid development. Consider how notoriously hard algorithms are to understand for policymakers. For them, explaining use toward a multiplicity of stakeholders is harder still. Choices, limitations, and tradeoffs reflected in the source code of an algorithm might be lost on the user of the insights produced by the algorithm, and those tradeoffs are only really made at the point of implementation of AI in a specific policy context for a specific public service or task (Baesens et al., 2015). Worse, if the policy maker does not understand the choices, and the limitations and caveats behind the insights, these may find their way into the decisions or policies supported by the AI (Van der Voort et al., 2019). Despite this, there are important lessons to take from each approach and opportunities to enhance the effectiveness of existing governance instruments by envisaging transparency of AI in more comprehensive terms that encompass multiple dimensions of the transparency problem. An integral approach is, after all, only a proportionate way to address a future challenge of technology that already implicates multiple actors and organizations in an integrally complicated way.

We contend that the different approaches are not to be viewed by scholars as contradictory or antagonistic to one another, but rather as different "realms" of transparency in the machinery of AI governance. As shown in Figure 31.1, four realms of AI transparency—legal (legal approach), organizational (capacity-building approach), political (democratization approach), and social (constructivist approach)—correspond to the four approaches to transparency of AI. While they take different approaches and contain valid points of disagreement, they should be seen not as rivals per se but approaches to different realms of governance affecting poles on two axes: (1) an *actor* axis varying between the transparency to citizens and transparency to domain experts such as administrators who use AI, and (2) a *structural* axis varying between the institutional and individual poles.

The political realm corresponds to the democratic perspective on transparency where the collective entity of the polity has a role supporting, maintaining, and realizing the institutional arrangements for transparency. The next realm is the legal realm, where legal precepts founded in the political system are developed that give statutory, enforceable power to expectations of transparency. The organizational realm corresponds to the capacity-building perspective because it concerns how resources, skills, and infrastructure possessed by organizations support transparency arrangements. The organizational realm can also be viewed as institutionally consequential to the development of transparency in the other realms of the political system and establishment of laws which are then shaped by capacity through implementation. Finally, the social realm corresponds to the constructivist perspective on transparency. Here the crux of transparency is human understanding and communication. Transparency is achieved through human relationships and shaped by how it is perceived and constructed through socio-technical processes.

We argue that all four realms should be considered in decision making whenever governments develop and/or implement AI systems. Decision makers should identify to whom they owe what form of transparency—experts or citizens—and at what level—institutional or individual—and use these different conceptions of transparency to consider when factoring transparency into the design of a system. Table 24.2 lists examples of the kinds of transparency tools that could be used in the smart meter example presented earlier (Table 24.1). This is not a requirements checklist that need be exhaustively or even progressively adopted but rather a set of illustrative examples according to each realm that decision makers can consider when tailor making their own transparency strategies.

Each transparency realm includes different kinds of tools, which could be accomplished individually or simultaneously so as to integrate strengths from each approach or to remedy

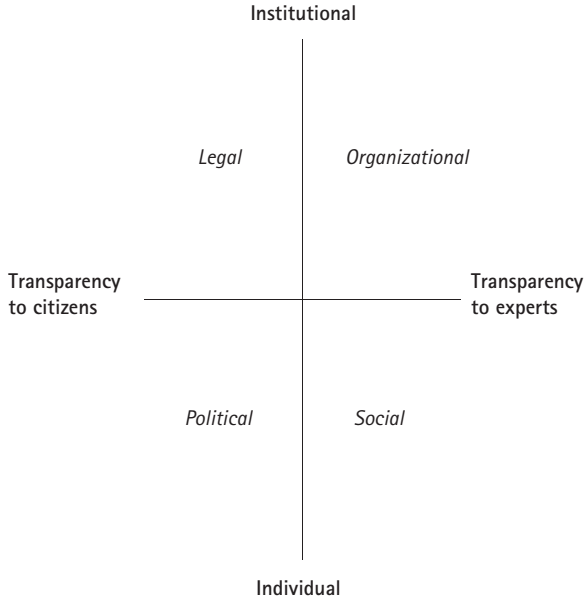

FIGURE 24.1 Realms of transparency in the design of AI governance

Table 24.2 Realms of transparent AI in smart meter design

Transparency Realm	Smart Meter Transparency Design
Political	• Public participation procedures for identifying what data should be shared and what kept private • Constitutional procedures for political and civil rights and complaints therein • Transparency agendas in democratic procedures such as elections, referenda, and legislative procedures
Legal	• Laws for open data and open-source AI programming of meters • Laws for explainability and interpretability of algorithms • Privacy laws that pertain to smart meter data • Procedural and statutory laws supporting the AI service architecture and environment • Laws around disclosure and contracting with smart meter companies
Organizational	• Computer knowledge and AI skills • Organizational protocols for AI decision making • Strategic planning and targets for AI programs • Employment and retention of specialist staff in smart technology, energy, and utilities • IT governance through oversight structures

Social	• Collaborative organizational design procedures with data specialists, energy companies, and local governments
	  • Empowerment of AI users
	  • Cultures of trust and accountability
	• Relational capabilities (between AI provision and use)

weaknesses of any realm in existing transparency architecture. As shown in the table, designers of transparent smart meter AI could choose among a series of tools that address the specific challenges faced by this type of AI. When the systems are built and provided by companies, transparency to the public may be driven by legal requirements on the procurement process, on energy performance, and on privacy. At the same time, obfuscation of the AI algorithms may occur as energy companies or regulators are incentivized to show performance in the best light possible. This requires that regulators, or other public sector professionals may understand what role algorithms play and how. At the level of a public organization this requires that they organize the capability to this end, which could be at the level of contracts, organizing capability externally (e.g., expert auditors) or training of personnel. If critical algorithmic information is hidden, learning processes for AI–human understanding will be harmed. Transparency is then only serviced when AI experts and non-experts understand what the other understands of their domain, which brings us closer to the social domain, which is in tune with the constructivist approach. In the social domain, transparency tools focus on culture change, individual empowerment, and relationship and trust building between system users and providers.

There are a myriad number of ways that organizations could choose among the four domains, giving priority to some of the domains over others. However, our main recommendation to policymakers seeking to find solutions to the transparency problem of AI is

a recommendation of integrality. Some of the transparency approaches seem to be *condicio sine qua non*, such as the development of capabilities. Ultimately, the implementation of AI transparency policies speaks to capability as embedded in a process that public sector actors need to engage in when considering or deploying AI. Such a process may involve code reviews; bias checks on data; explicating choices made in data processing, algorithm choice and parameterization, presentation, and visualization of results/insights; explicating use in decision making or policy; organizing interaction between different roles and types of professionals; and making explicit tradeoffs and choices regarding prioritization. The other side of the integration coin is diversification. We often talk about AI and algorithms in general, but the tools and their characteristics vary wildly, as do the requirements that use cases put on them. In some cases, transparency means that the general public must be able to acquire information about algorithm use, yet in other cases it suffices if experts such as auditors understand. Making explicit choices (e.g., between algorithm performance and level of explainability) serves transparency directly and indirectly by factoring transparency needs into the selection and configuration of policy tools.

Note

1. This example develops a case used in Giest, S. (2020). Making energy personal: Policy coordination challenges in UK smart meter implementation. *Journal of Public Policy 40*(4), 553–572.

References

Agarwal, P. K. (2018). Public administration challenges in the world of AI and Bots. *Public Administration Review 78*(6), 917–921.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society 20*(3), 973–989.

Andrews, L. (2019). Public administration, public leadership and the construction of public value in the age of the algorithm and "big data". *Public Administration 97*(2), 296–310.

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Bozeman, B., & Sarewitz, D. (2005). Public values and public failure in US science policy. *Science and Public Policy 32*(2), 119–136.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society 3*(1), 1–12.

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy 11*(5), 545–556.

Cofone, I. N. (2019). Algorithmic discrimination is an information problem. *Hastings LJ 70*, 1389.

Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ 105*, 1147.

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Admin. L. Rev. 71*, 1.

Criado, J. I., Valero, J., & Villodre, J. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity 25*(4), 449–470.

Dauvergne, P. (2020). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. *Review of International Political Economy 29*(3), 696–718.

Firlej, M., & Taeihagh, A. (2021). Regulating human control over autonomous systems. *Regulation and Governance 15*(4), 1071–1091.

Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & Governance 16*(1), 156–176.

Giest, S. (2020). Making energy personal: Policy coordination challenges in UK smart meter implementation. *Journal of Public Policy 40*(4), 553–572.

Gillis, T. B., & Spiess, J. L. (2019). Big data and discrimination. *The University of Chicago Law Review 86*(2), 459–488.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*. https://doi.org/10.1177%2F2053951719897945.

Harrison, T. M., & Luna-Reyes, L. F. (2022). Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review 40*(2), 494–511.

Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sciences 54*, 269–287.

Haunss, S., Kuhn, J., Padó, S., Blessing, A., Blokker, N., Dayanik, E., & Lapesa, G. (2020). Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance 8*(2), 326–339.

Heald, D. (2012). Why is transparency about public expenditure so elusive? *International Review of Administrative Sciences 78*(1), 30–49.

Henman, P. (2020). Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration 42*(4), 209–221.

Hood, C., & Heald, D. (2006). *Transparency: The key to better governance?* Oxford University Press for The British Academy.

Ingrams, A. (2019). Big data and Dahl's challenge of democratic governance. *Review of Policy Research 36*(3), 357–377.

Ingrams, A. (2020). A machine learning approach to open public comments for policymaking. *Information Polity 25*(4), 433–448.

Ingrams, A., Piotrowski, S., & Berliner, D. (2020). Learning from our mistakes: Public management reform and the hope of open government. *Perspectives on Public Management and Governance 3*(4), 257–272.

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly 33*(3), 371–377.

Klievink, B., Romijn, B. J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers 19*(2), 267–283.

König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly 37*(3), 101489.

König, P. D., & Wenzelburger, G. (2022). Between technochauvinism and human-centrism: Can algorithms improve decision-making in democratic politics? *European Political Science 21*, 132–149.

Ku, C. H., & Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly 31*(4), 534–544.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Matheus, R., Janssen, M., & Maheshwari, D. (2020). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly 37*(3), 101284.

McGregor, L. (2018). Accountability for governance choices in artificial intelligence: Afterword to Eyal Benvenisti's foreword. *European Journal of International Law 29*(4), 1079–1085.

Nicholson Price II, W. (2017). Regulating black-box medicine. *Mich. L. Rev. 116*, 421.

Panagiotopoulos, P., Klievink, B., & Cordella, A. (2019). Public value creation in digital government. *Government Information Quarterly 36*(4), 101421.

Sætra, H. S. (2020). A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society 62*, 101283.

Strandburg, K. J. (2019). Rulemaking and inscrutable automated decision tools. *Columbia Law Review 119*(7), 1851–1886.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly 36*(2), 368–383.

Van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making ? *Government Information Quarterly 36*(1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011.

Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart technology and the emergence of algorithmic bureaucracy: Artificial intelligence in UK local authorities. *Public Administration Review 80*(6), 946–961.

Williamson, B. (2014). Knowing public services: Cross-sector intermediaries and algorithmic governance in public sector reform. *Public Policy and Administration 29*(4), 292–312.

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review 21*(7), 1076–1100.

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., & Faraj, S. (2007). Information technology and the changing fabric of organization. *Organization Science 18*(5), 749–762.

CHAPTER 25

THE ANATOMY OF AI AUDITS

Form, Process, and Consequences

INIOLUWA DEBORAH RAJI

Introduction

When AI falls short of assumed or articulated expectations for performance, serious harm is almost certain to follow. Some may lose access to housing[1] or healthcare;[2] get falsely accused of a crime such as fraud;[3] or be unfairly rejected for a loan,[4] school admissions,[5] or a much-needed job.[6] The consequences can reverberate for months or even years, impacting any corner of one's life and livelihood.

In response to these concerns, recent years have seen an influx in the development and execution of AI audits—namely, evaluations of AI systems geared toward the goal of holding stakeholders accountable for such failures and their consequences. Published internal assessments by technology companies such as Facebook,[7] Google,[8] Microsoft,[9] and more[10] reveal the role of such audits in orienting corporations toward their own internal commitments to users, regulators, and self-defined AI principles. However, such *internal audits* embody perspectives that are limited in scope, failing to capture a more complete understanding of outside concerns. Conversely, *external audits*, executed by a range of stakeholders completely independent of the audited organization, can provide a novel set of expectations outside of corporate priorities and introduce critical considerations to challenge institutional narratives about how well an AI system works. These external auditors include those in investigative journalism,[11] civil society,[12] law,[13] government,[14] and academia.[15]

At the core of the objectives of AI auditing is the desire to reveal the quantitative and qualitative limits to the model as it relates to the concerns of those the auditor represents. This stakeholder-anchored assessment follows in the tradition of auditing practice in program evaluation,[16] environmental impact assessment,[17] finance reporting,[18] and more, revealing the value in soliciting a range of independent perspectives in determining the risks and benefits of an intervention or deployed artifact. The effectiveness of these audits is thus not just a question of reporting consistency in evaluation but an actual *design task* involving

reflection on what the objectives of an ideal model involve from the perspective of the auditor, and what it means to hold the developers of these AI systems accountable to those expectations. This requires conscious decision-making on how to best represent these desired outcomes in evaluation metrics, data, and methodology.

In this chapter, I first *define* an AI audit, differentiating it from related AI evaluation tools, such as algorithmic impact assessments and inspections. I describe the audit as embodying the twin goals of evaluation and accountability and discuss the difficulties of achieving such goals in the context of analyzing AI systems. Next, I break down the main categories of AI auditors—that of an *internal* or *external* auditor—and discuss the differences in motive, methodology, and constraints. Finally, I analyze the details of several case studies to discuss the key components and considerations of the audit process for each of these two types of auditors, highlighting how such audits can contribute to an entire ecosystem of AI governance measures, and result in meaningful steps toward accountability.

The Ongoing Challenge of AI Audits

In general, an audit lies at the intersection of evaluation and accountability—in fact, it can be defined as the evaluation of a system, executed for the purpose of accountability. Unlike an inspection or other form of assessment, with audits there is an expectation of access to the information required to make a qualified judgment with consequences. The goal of the audit is tangible impact—an update to standards, an engineered improvement, a legal challenge, the removal of a product from the market, or other intervention.[19] It is thus not just a designed assessment of formal or informal performance expectations, but one grounded in practical outcomes, often through informing updates to corporate behavior and the regulatory environment.

Vecchione et al. note the origin of audit studies in the social sciences from "activist research" intent on collecting empirical evidence of discrimination. These audits were thus "developed with direct participation of the affected communities and oriented around accountability and reform."[20] Inspired by these social science audits, Human–Computer Interface (HCI) researcher Christian Sandvig expresses a similar goal for the Internet platform audits he pioneered—designed as "public interest scrutiny of algorithms" to be "used as evidence in lawsuits alleging employment or housing discrimination."[21] Discussing the algorithmic audit in investigative journalism, Nicholas Diakopoulos re-iterates the role of audit work in exposing and examining "algorithmic power," again to hold decision-makers accountable for the consequences of their algorithmic deployments.[22] In the AI context, this aspiration has involved into the ideal of an "actionable audit,"[23] a phrase emphasizing the expectation of consequences and impact. Some also describe AI audit work as playing a "diagnostic" and "rebuttal" role in society, to further define the boundaries of acceptable AI use, and inform more compassionate, conscious mechanism design for future AI products.[24] Thus, the role of stakeholder perspective and collective evidence toward accountability outcomes has always been a central consideration in AI audit practice.

The details of what an AI audit looks like can vary, depending on context and the nature of the analyzed system. At times an AI audit can function as a binary judgment on certain process requirements or system outputs;[25] at other times, the audit looks like a very holistic

process of information seeking and judgment formation, with consequential though qualitative conclusions.[26] Given its origin in field experiments, many audit methods involve the active or passive participation of real users, who either donate their data or have their actions observed and then analyzed.[27] With computational systems, there is also the additional opportunity of analyzing *mock* users and leveraging independently curated benchmark data points to test for the performance of the system on a range of potential inputs in order to estimate the model's learned behavior.[28] Similarly, there are also opportunities for other strategies to be used, such as reverse engineering or simulation.[29] In the same vein, several strategies are employed by auditors to achieve their objective of tangible influence on audit targets.[30] Historically, the audit process has resulted in the accumulation of some form of evidence regarding a discriminatory or otherwise sub-par outcome. This evidence can then be leveraged as part of some formal accountability process (e.g., a lawsuit) or promoted through media or public advocacy to raise awareness of a concrete issue. This often leads to direct product updates or practical changes in corporate behavior and engineering practice, as well as modifications to the broader regulatory environment.

There is no straightforward definition for what could be classified as an "artificial intelligence" (AI) product. Some describe AI systems as involving any level of autonomous decision-making or computed demonstration of humanlike cognitive ability. Others speak specifically to the particular method of machine learning (ML), involving the processing of data to get to meaningful model predictions. In this chapter, I focus on the latter case, which seems to be the increasingly common form AI systems take. In this context, the primary distinction between AI tools and regular software tools is *data*. In regular software, the system output is set quite deterministically by the software engineer, who has fairly direct input and traceable influence on program outputs. With AI tools and, specifically, the common modern approach of ML, the system output is learned through the analysis of what is called a training dataset, and thus implicitly defined by the construction and processing of this dataset, rather than any explicit instructions in code.

Many of the practical challenges of designing an effective AI audit stem from the difficulty of executing evaluations and operationalizing accountability in the context of AI. There are certain properties of the technology that make each of these sub-goals taxing. In the following section, I discuss the unique challenges to executing both an effective evaluation practice, and meaningful accountability measure in the context of AI deployments.

Evaluation challenges in AI

An evaluation is a measurement of the gap between the articulated expectations of a system and its actualized performance.[31] In this case, it is the distance between what is expected, implicitly or explicitly, from AI systems and what is actually recorded in terms of product effectiveness or model behavior. AI systems, in particular, possess certain qualities that make its evaluation process even more unpredictable, inconsistent, and difficult. These challenges are mainly that of defining a *unifying standard*, dealing with *system opacity*, and handling *scale* or *context*.

As mentioned earlier, the goal of any evaluation is to express the discrepancy between the reality of a particular system's performance and the expectations articulated about that system. A *standard* is really a formalization of stakeholder expectations, and what a

stakeholder conducting the evaluation expects the product to achieve to confidently declare that a system "works." Sometimes the standard is not necessarily a benchmark on which the system is meant to achieve a certain performance, but also a reporting or procedural requirement. Standards and claims to performance can be initially proposed by those developing the system but can also involve the imposed expectations of other stakeholders.

In the AI industry, there are no clear standards for model development and use.[32] Part of this is due to the vague nature of what constitutes an AI system in the first place—these products are not simple to identify and are thus even more difficult to verify. The notion of "AI products" can encompass anything from application-specific models to broader general-purpose Application Program Interface (API) tools. This ambiguity in definition lends itself to exaggeration in performance claims, making it easier to add or remove any given technology from the categorization of AI to adjust expectations for product vetting. In fact, Mittelstadt cites the lack of consistency in the development and deployment practices of the AI industry as primary bottlenecks of progress for the field.[33]

Also, the metrics and benchmarks used to assess AI systems are far from unified or coherent.[34] In general, the AI field's understanding of the relationship between the measured performance on these benchmarks and real-world model behavior is quite shallow.[35] Beyond concerns for functionality and bias, there exist a diverse set of considerations to evaluate for, depending on the priorities of the involved stakeholders.[36] Often, there is a focus on quantitative results, but such expectations can also be expressed and judged qualitatively.

Another major challenge to evaluation in the AI context involves navigating the habitual *opacity* of such systems. To some degree, AI systems are usually self-deterministic, meaning that the model outcomes are mainly defined by the patterns in the data used in their development, and not the conscious decision-making of a programmer. As these datasets are often quite large—on the order of millions of images and text samples, sourced from various surveillance apparatus or the Web—it can be difficult to consistently and truthfully measure what the model is doing and how this relates to the engineering decisions informing its automated decision-making. Rarely will a reliable AI evaluation involve the scrutiny of the internal logic of a model's decision-making. The model itself operates as a mostly black-box system, making it difficult to judge the quality of the prediction. Despite some effort in this area, the ability to glimpse into the mechanisms through which predictions are made is ultimately still lacking.[37] In the absence of a clear view of internal processes, there is a tendency to over-rely on the outcomes of the model to assess its quality. AI evaluations are thus disproportionately focused on an analysis of such outcomes, often through a sample test dataset (i.e., "benchmark") used to judge the correctness of prediction outputs.

The final AI-specific evaluation challenge to discuss is the issue of managing performance expectations in the presence of *scale and context*. For a subset of AI tools—data-defined machine learning tools, in particular—it can be difficult to articulate expectations for a model because many models are developed from data of a particular context, but can also be easily distributed at scale. This means a single deployed model could possibly be canvassing a variety of contexts, some of which may be inadvertently incompatible. Effective AI evaluation thus also often needs to be quite context specific, with images representing specific cultures or demographics and text sourced from a particular dialect relevant to the scenario of deployment.[38] However, practitioners and researchers tend to not explicitly acknowledge this, and will usually over-claim model generalization, regularly implying that model performance metrics reported on a benchmark representing one specific context apply to

interpreting model performance in another situation or for a different population.[39] Such claims regularly disguise model failures that are context-specific.

As with all evaluations, the audit compares the reality of a system's operation with expectations of performance, as explicitly or implicitly communicated by any stakeholder. Practically, AI audits can also serve as a mechanism to articulate concrete concerns, in addition to establishing a minimum bar of performance. As there are several interpretations of what it means for an AI system to function, the audit can also sometimes play a clarifying role, providing concrete evidence of the current perceived state of the targeted AI product from the perspective of certain stakeholders.

Accountability challenges in AI

Similarly, accountability is another concept that can be difficult to pin down in the context of AI. Accountability is all about power and the ability to hold stakeholders in a position of power responsible for the impact they have on others. According to Wieringa,[40] AI accountability is tied to broader notions of an ability to hold those with influence on a situation responsible for their impact on the lives of those affected. It requires those with agency in a scenario—designated as "actors"—to provide enough information for those impacted (or their representatives)—named the "forum"—so that these impacted stakeholders can be able to make a judgment on the actor's decision-making. The forum's judgment should be consequential for the actor as a method of establishing checks and balances on the power the actor already holds in the given scenario (see Figure 25.1). If the goal is an accountability process that leads to different outcomes, there needs to be some improvement to our ability to enforce consequences in response to unfavorable audit results.[41] Here, I discuss the difficulty of achieving accountability outcomes in the context of AI systems, and particularly present the unique AI-related challenges of defining the stakeholder roles of actor and forum, as well as setting realistic accountability measures and enforcing them.

An algorithm cannot be punished for its decisions as an "actor" in typical accountability scenarios—it is thus unclear who is to be held responsible when an automated decision goes wrong. There are several stakeholders involved in shaping an AI system—from developers to model operators to data annotators—and many possess the power to operate as an actor in a given scenario involving AI systems. The data-defined models common in the AI industry still involve engineering decision-making (e.g., to determine the data source, define the label taxonomy, curate the dataset, etc.). However, in the AI context, those engineering decisions are much more difficult to track down, report and hold anyone responsible for. Thus, the relationship between engineering decision-making and system outputs are much less clear than in the software case and, as a result, it can be easy for those working on ML systems to deny responsibility, despite the reality that several of these stakeholders do determine, or at minimum contribute to, a model's capacity to do harm. The uncertainty around model ownership and influence makes it difficult to legally challenge algorithmic decision-making with harmful consequences.[42]

There is also ambiguity in the AI context regarding who exactly constitutes the "forum." Upon the deployment of an AI system, there are a range of impacted parties, represented by all kinds of institutions, that could constitute the forum in this scenario. AI systems tend to not only involve products that impact a direct user, but also these systems are more

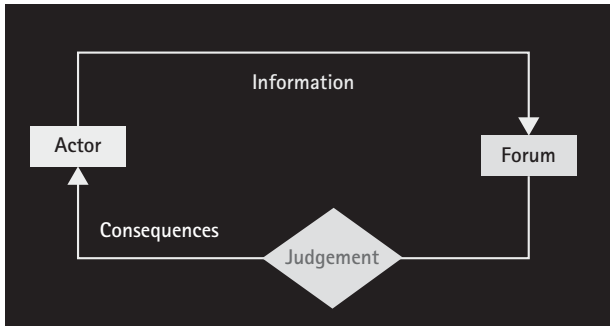

FIGURE 25.1 Accountability can be described as the relationship between an actor and a forum, the former being in a position to make decisions impacting the latter. An accountable system is one in which the actor has an obligation to explain and to justify their decision-making, and the forum can pose questions and pass judgment on the quality of those decisions, in such a way that the actor may face consequences in response to the forum's judgment.[44]

likely to influence life outcomes for a plethora of unknowing non-users as well. For instance, recruiters are the customers and users of automated hiring tools, even though such AI products impact job candidates the most. Similarly, doctors use medical AI tools on patients, teachers use automatic grading tools on students, and so on. As their awareness of the AI tool is almost completely dependent on the disclosure of its use, those ultimately impacted by AI systems can often not identify its influence on their lives.

Individuals included as data subjects, whose media (e.g., text or images) is informing a model, without due process for copyright or consent, are often also oblivious to their situation. As a result of this, and in consequence of the usual scale at which these systems are deployed, the "forum" is typically a large and untraceable crowd. Although some claim to operate as qualified representatives for the impacted population, many fall short, falling prey to faulty assumptions while mis-coordinating representation through the inappropriate grouping of other forum members. Nuance on how to solicit and navigate participation challenges is thus an integral consideration for achieving accountability.[43]

One of the most notable characteristics of an audit in the AI context is its required specificity. In general, AI audits need to scope evaluations to a particular target of interest and a specific performance claim if auditors wish to raise a meaningful challenge.[45] Given that AI tools are developed in often unreported contexts, the details of the scope of the scenario being assessed must be explicitly stated. This could involve audits aimed at a targeted product and, at times, even a targeted product feature or system capacity, inspecting claims that can range from vague promises of safety or fairness to specific assertions of achieving a certain accuracy measure in an implied or explicit range of contexts. Specificity in audit practice serves to allow for a clearer articulation of expectations and anticipated consequences.

This often differentiates AI audits from other algorithmic accountability tools such as algorithmic impact assessments (AIAs).[46] AIAs tend not to be as specific—rather than assessing AI performance in reference to specific claims, these are much broader risk assessment tools, often operationalized through more high-level thinking embedded in documentation guidelines. This is because the goal of AIAs is not as directly oriented toward the AI

audit goal of pushing to make judgments and garner consequences—the objective of AIAs is more to prompt reflection in AI system development and post-deployment monitoring to identify risks to potentially plan for, allowing for its more open-ended and flexible format.

Differentiating Between the Needs of Internal and External Auditors

Financial audit researcher Power mentions, "an audit is never purely neutral in its operations; it will operationalize accountability relations in distinctive ways."[47] Thus, the question of *who* conducts an audit is of ultimate importance. The auditor can represent the perspective of a range of external and internal stakeholders so, in general, I name two main types of algorithmic auditors—those "external" to the institutions developing the systems and those "internal" to these organizations.[48] Each type of auditor has their own distinctive interpretation of general audit goals. Internal auditors tend to design their analysis to protect against legal liability, quality control, compliance, and customer dissatisfaction. External auditors usually look at the AI system through the lens of the most at-risk populations, who are generally neglected and rendered invisible from typical calculations of accuracy, and they focus on accumulating the evidence necessary to raise awareness or advocate for some protective outcome.

One can think of internal auditors as those with contractual relationships with the audit target. This could mean internal audit teams at large corporations (known in ISO standards as "first-party" auditors*) or consultants operating within corporations (i.e., "second party" auditors). The practice of internal auditing is common in other disciplines, such as finance, medical device development, and aerospace.[49] The goal of internal auditors is to serve as a form of risk assurance and quality control, evaluating the performance of a system with respect to some internally mandated or externally imposed requirements. Internal audits are tied to notions of internal accountability, responsible innovation, and engineering responsibility—the objective is the measurement of compliance to inform internal product interventions in the engineering or business decision-making process, often prompted prior to the deployment of a product.[50]

On the other hand, external auditors (i.e., "third-party" auditors) are those with no contractual relationship to those developing the audited system. As a result, external auditors tend to have the ability to operate more independently, without any active consideration for the potentially conflicting interests of the audit target. However, as they have no formalized relationship to the audit target, they also have less access to the details of the system and less visibility into the functionality of that system, beyond often consumer-level exposure. These "external" auditors can be anything from regulators to advocacy groups, to legal firms or even investigative journalists—there is now a growing community of organizations and individuals representing outside perspectives to scrutinize these systems externally. The

* The International Organization for Standardization (ISO) defines first-party audits as internal audits; second-party audits as those performed by suppliers, customers, or contractors; and third-party audits as those performed by an independent body against a recognized standard.

goal of external auditors is to generate accountability from the outside, to assess corporate narratives of performance from the lens of those they represent, and to determine how well deployments operate for the communities these auditors represent.[51] This could practically mean collecting evidence for a legal trial, concretely articulating product limitations in a way that makes it legible to policymakers, or demonstrating the reality of concerns to inform public campaigns. These audits have as an objective to be "actionable," that is, to shift material realities for those impacted either through regulatory action, public dialogue, or a change in corporate behavior.[52] Especially in the context of advocacy, the legal compatibility of the involved analysis is necessary—measurements of disparity in AI performance for example, needs to be compatible with the local anti-discrimination law in order to serve as evidence that can be considered in court.[53]

A summary of the key differences between internal and external auditor characteristics can be found in Table 25.1. Examples of currently active internal AI auditors include, but are not limited to, the following: Internal audit teams, Ethical AI teams, or Responsible Innovation teams at major technology companies, such as IBM, Google, and Microsoft; large consultancies with an AI audit team, such as Deloitte, Accenture Responsible AI, and McKinsey & Company Quantum Black; and specialized consultancies, such as O'Neil Risk Consulting & Algorithmic Auditing (ORCAA). External auditors currently active in AI auditing include, but are not limited to, the following: regulators, such as the National Institute of Standards and Technology (NIST), the U.S. Food and Drug Administration (FDA), UK's Information Commissioner's Office (ICO), and more; investigative journalism groups, such as The Markup and Propublica; advocacy groups, such as the Algorithmic Justice League (AJL) or American Civil Liberties Union (ACLU); independent government or legal consultancies, such as Upturn; and specialized law firms, such as Foxglove.

Audits can play various roles throughout the design, development, and deployment of AI systems. Different stakeholders have different standards, according to their purpose for a

Table 25.1 A comparison of the tendencies of internal[54] and external[55] audit characteristics

Audit Characteristics	Internal Audits	External Audits
Involved Personnel	Executed by employees or contractors	Executed by independent, third parties
Objective	Focus on corporate compliance, quality control or liability	Focus on protecting represented groups, including non-users
Stage of execution	Pre-deployment	Post-deployment
Access Provided	Direct access to the system	Indirect and/or consumer-level access to the system
Standardization	Consistent format, and scheduled delivery	Ad-hoc, reactive, customized implementation
Audit Visibility	Direct communication and reporting to audit target	Reliance on public pressure and advocacy
Recommendation Enforcement & Impact	Reliance on internal cooperation	Reliance on legal, and advocacy levers

particular evaluation. For example, external audits can serve as part of regulatory standard development procedures and a procurement process to determine vendor eligibility for government contracts could involve a required internal audit executed by vendors as part of the vetting process. External audits also sometimes feed into advocacy and campaigning outside of regulator action to raise awareness or update public opinion regarding certain harmful AI failures. Internal audits, on the other hand, tend to play a role in product performance assessment as part of the product development cycle to set deployment criteria for widespread or even pilot release. These kinds of audits can also guide sales restrictions or recalls (voluntary or not). Thus, audit outcomes could prevent products that do not hit a minimum acceptable bar of performance from proliferating or reaching the market in the first place. This raises the possibility of audits factoring into strategies to combat product liability and consumer protection concerns, thus protecting the public from being sold products that are unsafe or falsely advertised. Conversely, internal auditors can also leverage audit results to serve marketing and promotional purposes to find ways to communicate superior product performance when compared to competitors.

Access

The primary difference between internal and external audit contexts is *access*. This constrains many external auditors, forcing them to conduct their evaluations differently. Internal audit strategies tend to be discursive and heavily documentation based. This involves everything from engaging in procedural interviews and mapping internal stakeholders[56] to setting up more robust model reporting[57] or data documentation.[58] These documentation efforts are crucial interventions—providing a record for broader communication of the state of a system to a wider range of stakeholders and making legible failure modes and engineering design decisions that otherwise would remain unrecorded or not executed. As many documentation guidelines require the cooperation and visibility of an engineering process closed off from external auditors, such an approach for accountability and evaluation is simply not available to external auditors.

Interestingly, much of the currently proposed algorithmic accountability policy is geared toward an assumption of access and so disproportionately accommodates the needs of internal, rather than external auditors. For example, the UK's ICO released guidelines for auditability that required direct engagement with the involved engineering team to fulfill documentation requirements.[59] Similarly, the United States 2019 Algorithmic Accountability Act relies on corporate cooperation in order to fulfill reporting requirements.

A key aspect of accountability is the ability to demand the information required to make a valid judgment—it is clear that external auditors currently do not have such guarantees. In fact, several external auditors have already been accused of breeches to company Terms of Service agreements, the Computer Fraud and Abuse Act (CFAA) or the Association for Computing Machinery (ACM) ethical code of conduct in their desperate attempts to access the data required to execute their evaluations.[60] It is only quite recently that policymakers have begun to include external auditors in policy dialogue, and to consider the need to properly mandate companies to provide access to those providing outside scrutiny. That being said, the acknowledgment is incomplete and non-inclusive. American developments, such as the Algorithmic Justice and Online Platform Transparency Act, the California

Algorithmic Accountability Bill, and a 2021 re-write of the Algorithmic Accountability Act, include some consideration for auditors coming from outside perspectives, although the focus seems to be on empowering third-party regulator auditors almost exclusively. Relatedly, the EU Commission's Digital Services Act (Articles 28 and 31) mentions the need to provide access to outside scrutiny, but oddly single out only academic researchers as valid external auditors, in the process neglecting advocates, legal firms, investigative journalists, and more that presently occupy this role.

Standardization

In the same way, internal audits can be more easily *standardized*. Although the variability of the current AI development process makes it difficult to assess ethical or functional compliance,[61] there is a rough consistency to key events in the development cycle and broad corporate structures that internal audit procedures can anchor themselves to. This allows for the availability of internal audit templates and consistent comparisons across audit target outcomes.[62] External audits represent the specified concerns of any given external community—there is thus a high variability in the types of concerns brought forth and assessed during the audit process. Even standard measures like accuracy evolve depending on the context of who the model is being evaluated for and the perspective through which notions of safety, fairness, and even privacy are defined. On one hand, that makes each external audit unique to its own context and questions, rarely involving too much overlap with past inquiries. This makes it even more challenging to develop consistent methodology, templates, or tools, but it also provides an opportunity for a richer understanding of an AI system's performance. For instance, many external auditors embedded at regulatory bodies[63] will evaluate tools completely differently from advocacy groups,[64] who will differ from the analysis of investigative journalists.[65] Each will adapt their strategy and findings to their own context, capabilities, and priorities, providing a more complete assessment of the AI tool than if the analysis was to rely on a single audit performance.

Audit visibility and impact

Finally, both types of auditors have clear differences in their *visibility* to the audit target and the public. In general, it can be difficult for external auditors to be taken seriously. Internal auditors have direct contact with internal stakeholders—as employees or contractors, they can more easily command the attention of relevant actors within the company to enforce whatever consequences or interventions are necessary in response to an audit result. External auditors, however, can be easily dismissed by corporations—ignored by corporate stakeholders and often with no direct access to demand accountability. Many early online audits were ignored by audit targets for this very reason.[66] It was not until recently that a resurgence of audits actively designed to yield actionable outcomes has succeeded in capturing corporate attention through strategies like leveraging public pressure and naming multiple targets.[67] On a related note, disclosure practices greatly vary between audit types—external auditors, who are not contractually bound and often approach the process with a public interest mandate, tend to release audit results publicly. Meanwhile, internal audit results are

rarely externally visible because internal auditors are likely subject to non-disclosure terms as part of their contract with the involved corporations.

Case Studies of AI Audits

All in all, there are several open challenges in how audits are designed and executed in the AI context. Many of these ongoing questions revolve around the less developed external audits, although all AI audits are still very much works in progress, evolving as increasingly effective algorithmic accountability tools. The difference between both types of auditors and their corresponding strategies manifests in certain important methodological differences between audits coming from outside stakeholders and those coming from the inside, each with their own strengths and weaknesses. For example, those on the inside definitely have certain advantages, such as increased access and visibility, but obtaining a diversity of perspectives and maintaining audit integrity is easier to achieve as an external auditor, independent of corporate influence and clear of explicit conflict of interest.

In this section, I analyze the details of the dilemmas involved in AI audit design, through an exploration of two primary case studies—one internal audit and one external audit—highlighting the advantages and difficulties of both approaches in supporting downstream accountability outcomes. I then conclude with some complementary examples from a range of other domains to demonstrate how, although the goals of AI auditing are consistently anchored to expectations of accountability, auditor behaviors and audit design regularly influences the outcomes of the process.

Internal audits and AI hiring tools

Many have long acknowledged the need for the further scrutiny of AI's deployment as part of hiring processes. Nearly all Fortune 500 companies now use algorithmic recruitment and hiring tools.[68] Experts such as Professor Ifeoma Ajunwa have long sounded the alarm. In her work, she presents AI audits as an essential intervention and accountability mechanism for the industry, highlighting the approach as an opportunity to critically examine product performance in light of tangible employment discrimination threats.[69]

In general, internal audit approaches are anchored to process-level interventions to slow down the pace of development and prompt further developer reflection. This is why internal auditors tend to depend so heavily on approaches such as documentation strategies[70] and checklists[71] (i.e., tools for those building the system to record and verify the details of their decision-making). The objective of these internal methods is to operationalize the internal expectations of the target organization and set criteria for deployments that are consistent with the organization's own guiding AI principles and legal obligations.[72]

Much like education, credit, and housing discrimination, employment discrimination falls into a domain under U.S. legal jurisdiction and thus the *legal compatibility* of the audit details needs to be considered. In fact, matching audit metrics to the existing anti-discrimination law will determine the effectiveness of the internal audit in its goal of protecting the audited organization from downstream liability.[73]

Raghavan et al. did a survey on AI hiring tools, and their publicly reported internal audit outcomes, visible in a limited way through press releases, marketing materials, and the occasional publication. What this survey revealed is that the companies were almost universally preoccupied with fairness issues.[74] Of notable concern was the "four-fifths rule" for employment discrimination in the United States—namely, that "if the selection rate for one protected group is less than 4/5 of that of the group with the highest selection rate, the employer may be at risk."[75] In recent detailed internal audits of the AI hiring products created by companies such as Pymetrics[76] and Hirevue,[77] the audit focus was also on the satisfaction of the "four-fifths rule" and fairness concerns. With full access to internal systems, both consultancies serving as hired internal auditors for Pymetrics and Hirevue could fully examine the product for compliance to the expectations of legal anti-discrimination and the related liabilities for their customers, the employers. In both cases, the examined products "passed the audit," demonstrating adherence to the four-fifths rule.

However, in the same survey, Raghavan et al. find that only five out of the 18 vendors reviewed made any form of validation claim in their marketing materials, with just two demonstrating evidence of executing studies on the tool's effectiveness.[78] Fitting this description, the Pymetrics and Hirevue studies make little explicit mention of any validation processes for their use of scientifically contested personality tests and facial recognition-based personality analysis to determine nebulous notions of "cultural fit" and worker compatibility.[79] In general, the focus on fairness in this case somewhat disguised the lack of basic functionality common in the industry and consequently distracted from the accompanying lack of validation for these tools before their entry into the market and mainstream use.

Both the Pymetrics and Hirevue audits were executed by paid consultants, under the legal binding of a Non-Disclosure Agreement (NDA). This led to heavy criticism and the accusation that these audits were incomplete in their scrutiny of these system.[80] In both cases, the involved companies published some form of public summary of main audit findings. However, each publication included clear oversight from the audited company—either in the form of corporate employee co-authors[81] or corporate-controlled restrictions to the dissemination of the final audit report.[82] As noted in literature,[83] a lack of independence from audit targets is a clearly visible threat to the perceived credibility of internal audits.

However, the benefits to internal auditing are also quite visible in this case. In both cases, internal auditors were successful in assessing legal liabilities and were also able to point out additional ethical and performance blind spots to the audited companies, providing some additional recommendations. The non-adversarial nature of the dynamic between internal auditor and audit target also clearly allowed for more opportunities for the meaningful enforcement of auditor recommendations, in addition to the much-needed access necessary to complete assessments. Framed as a "cooperative audit," the Pymetrics audit team encouraged as next steps in the auditing process to question "the efficacy of pymetrics' games at assessing the 'fit' of job seekers," including the possibility of making the current internal validation testing public.[84] In the case of Hirevue, the auditors brought into question the product's use of facial recognition, recommending that Hirevue drop the integration of that technology from their platform—which the company did, over a year later. Thus, it is clear that internal audits can serve their purpose in anchoring companies to a higher standard of performance relative to their admittedly narrow goals of legal compliance and internal consistency.

External audits and facial recognition

In 2018, the Gender Shades study[85] revealed the biased performance of commercial facial analysis products and spurred a remarkable corporate and public response. This audit provides a clear illustration of some of the challenges highlighted in earlier sections, revealing both the opportunity and limitations of external AI audit approaches.

Facial recognition technology (FRT) systems are one of the most controversial modern applications of AI. FRT systems can be weaponized against minority populations in various ways. Although early datasets would record the demographics and solicit the consent of photo-shoot subjects, this practice disappeared once web-based sources and larger datasets became the norm.[86] As a result, many of the canonical datasets in the field are dubiously sourced and heavily skewed in demographic representation, with the majority of subjects being of a lighter skin type, male, and middle aged.[87] This leads to performance differences between demographic subgroups that can put those groups at risk—for instance, misidentified Black and transgender drivers sued due to difficulties with ride sharing company Uber's facial recognition authentication process.[88] Similarly, the darker skinned individuals more likely to be misidentified by FRT models deployed in law enforcement cases are also those at higher risk of such identifications escalating to serious consequences, such as a false arrest.[89]

The authors of the Gender Shades study, Gebru and Buolamwini—both Black women—purposefully proposed an intersectional analysis across skin type and gender of the performance of deployed facial analysis API tools developed by tech giants like Microsoft, IBM, and the Chinese company Megvii.[90] Given the bias of mainstream facial recognition benchmarks at the time,[91] the authors cautiously sourced their own test set of public figures, intentionally balancing the representation of men and women, as well as darker skinned individuals and lighter skinned subjects. The involved evaluation calculations for the audit itself are quite simple—defined by the reporting of differences in subgroup classification error, they executed a standard form of disaggregated analysis, already common in other fields like healthcare.[92] The test set they developed (named the Pilot Parliaments Benchmark) was carefully crafted to optimize for the accuracy and visibility of the audit.[93] The authors crafted engaging demos and a website, investing energy in how to translate their audit results into visible artifacts that would resonate with a public audience.[94]

In addition to design decisions regarding the format of the audit, updates were made to the audit process. The audit targets themselves were intentionally selected and named. Rather than leaving corporate audit targets anonymous or focusing on a single target, the authors leveraged naming multiple audit targets to garner the press and public attention required to pressure companies to pay attention to audit results. Instead of relying on an audit of the end product, they audited the AI models directly through public APIs to address challenges of access as external auditors. Most importantly, to mitigate corporate hostility, auditors approached the audit targets before the audit's public release, adopting a coordinated disclosure policy similar to that observed for software bug reporting in information security.[95]

The result was worth it—the 2018 audit study revealed that Microsoft, IBM, and Megvii's *deployed* facial analysis products were up to 34.4 percent less accurate on darker-skinned female subjects than lighter-skinned male subjects.[96] A year later, in 2019, a follow-up study

revealed that all three companies released new API product versions within seven months of the publication of the first audit, shrinking performance differences between demographic groups. That being said, an audit of new products in the same industry revealed that there was still a long way to go. Amazon's Rekognition product, also an AI facial analysis tool, was still over 30 percent less effective on darker-skinned female faces than lighter-skinned male faces.[97] This audit was conducted on August 21, 2018—using the exact same product version that Amazon was at the time attempting to pitch for use by police departments and the U.S. Immigration and Customs Enforcement (ICE) office. By June 2020, influenced by public advocacy campaigns following the Gender Shades audit release, audit targets IBM, Microsoft, and Amazon had all made public corporate statements denouncing the use of the technology[98]—IBM, in particular, voluntarily pulled out of the facial recognition market altogether and Amazon set a one-year voluntary moratorium on the sale of its technology to police.[99] By May 2021, this moratorium was expanded indefinitely.[100]

More importantly, the Gender Shades audit opened a pandora's box for questioning the functionality of FRT products. Subsequently published pilot results demonstrated alarmingly high real-world error rates[101] (including one pilot with a full 100 percent error rate[102]). Follow-up audits by civil society organizations like the ACLU[103] sparked nationwide campaigns resulting in tangible policy reform, including several municipal and state-wide facial recognition bans.[104] Finally, results were replicated by the regulator NIST, which integrated a similar audit process to its 2019 Facial Recognition Vendor Test (FRVT).[105]

That being said, there were limitations to the audit, and these revealed themselves when analyzing the longer-term consequences of the audit's outcomes. As with most external AI audits, the Gender Shades audit was geared toward narrow impact. With tightly scoped evaluation results, the audit's impact would only mainly influence the performance of the named audited corporations, for the specific audited demographic, and the specified task being audited.[106] For example, the initial Gender Shades audit assessed the performance of these commercial systems for the facial analysis task of gender classification. This task was selected because it was simple (i.e., binary output), and connected to a shared experience of the co-authors (i.e., Black women who had been mis-gendered by FRT in the past). However, once released, the audited corporations seemed to focus their product updates on the gender classification task specifically, still maintaining performance disparities for other tasks, such as smile detection or age classification.[107] Given the controversial nature of automatic gender classification as a task,[108] and the range of other FRT tasks with disparate performance, this was a particularly disappointing audit response. Similarly, the definition of intersectional audited subgroups by gender and skin type disguised disproportionate product failures related to other demographic attributes, such as age. Also, even though the selected audit targets of Amazon, IBM, Microsoft, and more took action in response to the audit results, other less publicly visible FRT companies, such as NEC or Cognitec, did not feel the same pressure to address the raised problem.

Auditor responses were quite varied and revealed much about the inherent back-and-forth that follows the release of an external audit. For instance, audit target Amazon, who initially behaved defensively in response to unfavorable audit outcomes, would initially attempt to deny audit results. At one point, the company claimed that audit results were untenable because supposedly the default prediction threshold settings used in the audit differed greatly from their recommended settings for police clients. However, when interviewed, these clients were completely unaware of what those recommended thresholds were, revealing this

difference did not bear out in practice.[109] Similarly, the company refused to participate or tolerate other external audits conducted, such as the follow-up study from NIST.[110] Eventually, the company could not refute the plethora of reproduced results and defense of the study coming from the scientific community.[111] Similar resistance came from audit target Kiaros, who responded to being audited by increasing their paywall and updating to a more restrictive terms of service agreement in an attempt to keep future external auditors out.[112]

Even audit targets that co-operated with the external auditors and accepted audit results did not always respond as expected. For instance, IBM responded to Gender Shades with the development of the Diversity in Faces dataset, a dataset including a range of more diverse face data. This release contributed to minimizing performance differences across demographic groups[113] but was later revealed to include faces sourced from Flickr without consent.[114] Interestingly, Facebook, a company that could not be audited externally because their facial recognition model was not public, also responded to the audit by creating a new dataset of diverse faces, but this time sourced with consenting models.[115] Facebook also engaged in a voluntary, albeit incomplete, recall of features involving facial recognition.[116]

On-the-surface wins could not really be trusted. Microsoft's response to the Gender Shades audit involved a push for regulatory involvement in facial recognition, an investment that ultimately led to their control of the legislative process for FRT regulation in the state of Washington.[117] Companies such as Amazon, which committed to a moratorium on the sale of the technology to police, were able to set the terms of their own voluntary recall. As a result, they continued to sell the technology under a different form. In this case, Amazon continued offering facial recognition functionality to police through partnerships with their smart doorbell product, Ring—a client base that has grown from just 400 police departments at the time of the audit in 2019[118] to over 2,000 in 2021.[119] Similarly, the enforcement of the celebrated regulatory bans were rife with loopholes, which allowed for police departments in cities where the technology was banned to continue to send images for analysis remotely.[120]

Also, even in cases of low classification bias, it has become increasingly evident over the years that FRT can still be weaponized for discriminatory practices. Functionality can become meaningless in the context of an inherently invalid task, or completely inappropriate misuse. For example, some facial analysis tasks simply regurgitate the "discredited pseudosciences of physiognomy and phrenology,"[121] where a subject's inner state is wrongly assumed through the evaluation of that subject's external facial features (e.g., predicting criminality or sexuality from one's face). Also, there have been several instances of the technology being completely misused or weaponized against vulnerable groups, with police reportedly using incorrect image inputs, such as face sketches and celebrity photos, to generate false matches with the facial recognition technology,[122] and activists being regularly monitored through face surveillance.[123] Such inappropriate applications or uses do not necessitate an empirical audit.

Conclusion

AI audit work has been transformational for the industry. AI audits developed by investigative journalists at The Markup and ProPublica were able to publicly expose life-altering product failures in tenant screening,[124] criminal risk assessment,[125] insurance pricing,[126] and online advertising.[127] Academic audits revealed biased outcomes by algorithms used

for content curation,[128] child welfare,[129] criminal justice,[130] hiring,[131] and public health.[132] Civil society organizations like the ACLU, regulators like NIST, and tech-specialized law firms like Foxglove have leveraged AI audits to generate evidence to protect their clients in lawsuits from the Ofqual A-level grade adjustment algorithm,[133] the Home Office automated visa processing system,[134] and facial recognition.[135] And internal audits led to voluntary recalls and radical redesigns, getting companies to think twice before mass deployment.[136]

This impact is intentional, and a core aspect of what it means to conduct an effective audit. An algorithmic audit involves the collection and analysis of outcomes from a fixed algorithm or defined model within a system, with the goal of holding those that control and define that system accountable for its impact. Targeted external AI audits provide one mechanism to incentivize corporations to address the algorithmic issues present in the data-centric technologies that continue to play an integral role in daily life, from governing access to information and economic opportunities to influencing personal freedoms. Internal audit mechanisms ensure some level of due diligence during the engineering development process and accountability to a company's own declared principles. Regardless of the precise strategy involved, it is clear that a primary contribution of auditing is the ability to analyze AI systems from a fresh perspective and tie that analysis to demands for action. As we collectively work through the details of formalizing audit design processes and addressing limitations, the field is just now beginning to acknowledge the value of AI audits—and, in particular, the value of third-party AI audits—as meaningful mechanisms for accountability and AI governance.

Notes

1. Kirchner, L., & Goldstein, M. (2020a). Access denied: Faulty automated background checks freeze out renters. *The Markup*; Kirchner, L., & Goldstein, M. (2020b). How automated background checks freeze out renters. *The New York Times*, May 28.
2. Lecher, C. (2018). What happens when an algorithm cuts your health care. *The Verge*, March 21.
3. Charette, R. (2018). Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold-IEEE spectrum. *IEEE Spectrum 18*(3), 6; Hill, K. (2020). Wrongfully accused by an algorithm. *The New York Times*, June 24.
4. Martinez, E., & Kirchner, L. (2021). The secret bias hidden in mortgage-approval algorithms. *The Markup*.
5. Kippin, S., & Cairney, P. (2021). The COVID-19 exams fiasco across the UK: Four nations and two windows of opportunity. *British Politics 17.1*, 1–23.
6. Dastin, J. (2022) Amazon scraps secret AI recruiting tool that showed bias against women. Ethics of data and analytics. *Auerbach Publications*, 296–299.
7. Facebook. (2020). Indonesia human rights impact assessment. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Indonesia-HRIA.pdf.
8. BSR. (2019). Google celebrity recognition API human rights assessment. https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf; Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

9. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM 64*(12), 86–92; Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
10. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–16; Raji, I. D., & Yang, J. (2019). About ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv preprint arXiv:1912.06166*; Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human–Computer Interaction 5*(CSCW1), 1–23.
11. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May 23, 139–159; Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism 3*(3), 398–415; Kirchner & Goldstein, 2020a; Kofman, A., & Tobin, A. (2019). Facebook ads can still discriminate against women and older workers, despite a civil rights settlement. *ProPublica*, December 13; Varner, M., & Sankin, A. (2020). Suckers list: How AllState's secret auto insurance algorithm squeezes big spenders. *The Markup*.
12. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91; Lum, K., & Isaac, W. (2016). To predict and serve? *Significance 13*(5), 14–19; Snow, J. (2018). Amazon's face recognition falsely matched 28 members of congress with mugshots. https://www.aclu.org/blog/privacytechnology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.
13. Foxglove. (2020a). Home office says it will abandon its racist visa algorithm—after we sued them. https://office says it will abandon its racist visa algorithm after we sued them/; Foxglove. (2020b). We put a stop to the A level grading algorithm! https://www.foxglove.org.uk/2020/08/17/we-put-a-stop-to-the-a-level-grading-algorithm/.
14. Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (FVRT): Part 3, demographic effects*. National Institute of Standards Technology.
15. Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Conference on Fairness, Accountability and Transparency*, 134–148; Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453; Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM 56*(5), 44–54.
16. Posavac, E. J. (2015). *Program evaluation: Methods and case studies*. Routledge.
17. Hughes, R. (1998). *Environmental impact assessment and stakeholder involvement*. International Institute for Environment and Development. https://www.jstor.org/stable/resrep18000.
18. Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.
19. Ibid.
20. Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

21. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry 22*, 4349–4357.
22. Diakopoulos, 2015.
23. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.
24. Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260.
25. Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction 5*(CSCW1), 1–34; Sandvig et al., 2014.
26. Vecchione et al., 2021.
27. Angwin et al., 2016.
28. Sandvig et al., 2014.
29. Lum and Isaac, 2016.
30. Raji & Buolamwini, 2019.
31. Mathison, S. (2004). *Encyclopedia of evaluation*. SAGE.
32. Butcher, J., & Beridze, I. (2019). What is the state of artificial intelligence governance globally? *The RUSI Journal 164*(5–6), 88–96.
33. Mittelstadt, B. (2019b). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1*(11), 501–507.
34. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*; Verma, S., & Rubin, J. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, 1–7.
35. Liao, T., Taori, R., Raji, I. D., & Schmidt, L. (2021). Are we learning yet? A meta review of evaluation failures across machine learning. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
36. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*(9), 389–399.
37. Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.
38. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*.
39. Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.
40. Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18.
41. Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics 160*(4), 835–850.
42. Kroll, J. A. (2015). *Accountable algorithms* (Doctoral dissertation). Princeton University.
43. Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.
44. Ibid.

45. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.
46. Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746.
47. Power, 1997.
48. Raji, I. D., Costanza-Chock, S., & Buolamwini, J. (2022). Change from the outside: Towards credible third-party audits of AI systems. *Missing Links in AI Policy*.
49. Raji, Smart, et al., 2020.
50. Ibid.
51. Raji & Buolamwini, 2019.
52. Ibid.
53. Xiang, A. (2021). Reconciling legal and technical approaches to algorithmic bias. *Tennessee Law Review 88*(3), 649; Xiang, A., & Raji, I. D. (2019). On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*.
54. Raji, Smart, et al., 2020.
55. Raji & Buolamwini, 2019.
56. Raji, Smart, et al., 2020.
57. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
58. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
59. Kazim, E., Denny, D. M. T., & Koshiyama, A. (2021). AI auditing and impact assessment: According to the UK information commissioner's office. *AI and Ethics 1*, 1–10.
60. Kadri, T. (2020). Digital gatekeepers. *Texas Law Review 99*, 951; *Sandvig v. Barr*. (2020); *Van buren v. United States*. (2020).
61. Mittelstadt, B. (2019a). AI ethics—too principled to fail. *arXiv preprint arXiv:1906.06668*.
62. Raji, Smart, et al., 2020.
63. Grother et al., 2019.
64. Buolamwini & Gebru, 2018; Snow, 2018.
65. Diakopoulos, 2015.
66. Raji & Buolamwini, 2019.
67. Ibid.
68. Ajunwa, I. (2021). The auditing imperative for automated hiring. *Harvard Journal of Law & Technology*, *34*, 621.
69. Ibid.
70. Gebru et al., 2021; Mitchell et al., 2019; Raji, Smart, et al., 2020.
71. Madaio et al., 2020.
72. Holstein et al., 2019; Rakova et al., 2021.
73. Xiang & Raji, 2019.
74. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.
75. Greenberg, I. (1979). An analysis of the EEOCC "four-fifths" rule. *Management Science 25*(8), 762–769.

76. Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677.
77. O'Neil Risk Consulting & Algorithmic Auditing (ORCAA) (2020). Description of algorithmic audit: Pre-built assessments. https://techinquiry.org/HireVue-ORCAA.pdf.
78. Raghavan et al., 2020, p. 8.
79. Raghavan et al., 2020; Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3927300.
80. Engler, A. C. (2021). Independent auditors are struggling to hold AI companies accountable. *Fast Company*, January 26. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue.
81. Wilson et al., 2021.
82. ORCAA, 2020.
83. Raji, Smart, et al., 2020.
84. Wilson et al., 2021.
85. Buolamwini & Gebru, 2018.
86. Raji, I. D., & Fried, G. (2021). About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*.
87. Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in faces. *arXiv preprint arXiv:1901.10436*; Raji, Gebru, et al., 2020.
88. Melendez, S. (2018). Uber driver troubles raise concerns about transgender face recognition. *Fast Company*, August 9. https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers.
89. Hill, 2020.
90. Buolamwini & Gebru, 2018.
91. Merler et al., 2019.
92. Kauh, T. J., Read, J. G., & Scheitler, A. (2021). The critical role of racial/ethnic data disaggregation for health equity. *Population Research and Policy Review 40*(1), 1–7.
93. Raji & Buolamwini, 2019; Raji, Gebru, et al., 2020.
94. Raji & Buolamwini, 2019.
95. Ibid.
96. Buolamwini & Gebru, 2018.
97. Raji & Buolamwini, 2019.
98. Magid, L. (2020). IBM, Microsoft, and Amazon not letting police use their facial recognition technology. *Forbes*, June 12.
99. Hao, K. (2020). The two-year fight to stop Amazon from selling face recognition to the police. *MIT Technology Review*, June 12.
100. Dastin, J. (2021). Amazon extends moratorium on police use of facial recognition software. *Reuters*, May 18. https://www.reuters.com/technology/exclusiveamazonextendsmoratoriumpoliceusefacial-recognition-software-2021-05-18/.
101. Manthorpe, R., & Martin, A. J. (2019). 81% of "suspects" flagged by MET's police facial recognition technology innocent, independent report says. *Sky News*, July 4. https://news.sky.com/story/metpolices-facial-recognition-tech-has-81-error-rate-independent-report-says-11755941.
102. Berger, P. (2019). MTA's initial foray into facial recognition at high speed is a bust. *Wall Street Journal*, April 7. https://www.wsj.com/articles/mtas-initial-foray-into-facial-recognition-at-high-speed-is-a-bust-11554642000.

103. Snow, 2018.
104. ACLU. (2019). Community Control Over Police Surveillance. https://controloverpolicesurveillance; Haskins, C. (2019). Oakland Becomes Third U.S. City to Ban Facial Recognition. Vice, July 17. https://www.vice.com/en/article/zmpaex/oakland-becomes-third-us-city-to-ban-facial-recognition-xz.
105. Grother et al., 2019.
106. Raji, Gebru, et al., 2020.
107. Ibid.
108. Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human–Computer Interaction 2*(CSCW), 1–22.
109. Menegus, B. (2019). Defense of Amazon's face recognition tool undermined by its only known police client. https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149.
110. Harwell, D. (2019b). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *Washington Post*, December 19.
111. Bass, D. (2019). Amazon schooled on AI facial technology by Turing award winner. https://www.bloomberg.com/news/articles/2019-04-03/amazon-schooled-on-ai-facial-technology-by-turing-award-winner?in_source=embedded-checkout-banner.
112. Raji & Buolamwini, 2019.
113. Ibid.
114. Merler et al., 2019.
115. Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2021). Towards measuring fairness in AI: The casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science 4*(3), 324–332.
116. Heilweil, R. (2021). Facebook is backing away from facial recognition. meta isn't. *Vox*, November 3. https://www.vox.com/recode/22761598/facebook-facial-recognition-meta.
117. Simonite, T. (n.d.). Congress is eyeing face recognition, and companies want a say. WIRED. https://www.wired.com/story/congress-eyeing-face-recognition-companieswant-say.
118. Harwell, D. (2019a). Doorbell-camera firm Ring has partnered with 400 police forces, extending surveillance concerns. *Washington Post*, April 28.
119. Lyon, K. (2021). Amazon's Ring now reportedly partners with more than 2,000 us police and fire departments. *The Verge*, January 31.
120. Harwell, D. (2021). Police say they can use facial recognition, despite bans. *The Markup*. https://themarkup.org/news/2021/01/28/police-say-they-can-use-facial-recognition-despite-bans.
121. y Arcas, B. A., Mitchell, M., & Todorov, A. (2017). Physiognomy's new clothes. *Medium*, May 6. https://medium.com/@blaisea/physiognomys-new-clothesf2d4b59fdd6a.
122. Garvie, C. (2019). Garbage in, garbage out. face recognition on flawed data. *Georgetown Law Center on Privacy & Technology*.
123. Garvie, C. (2016). The perpetual line-up: Unregulated police face recognition in America. *Georgetown Law, Center on Privacy & Technology*.
124. Kirchner & Goldstein, 2020a.
125. Angwin et al., 2016.
126. Varner & Sankin, 2020.
127. Kofman & Tobin, 2019.
128. Sweeney, 2013.

129. Chouldechova et al., 2018.
130. Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior 46*(2), 185–209.
131. Raghavan et al., 2020.
132. Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., & Ho, D. E. (2021). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for covid-19 policy. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 173–184; Obermeyer et al., 2019.
133. Foxglove, 2020b.
134. Foxglove, 2020a.
135. Grother et al., 2019; Snow, 2018.
136. Chouldechova et al., 2018; Knight, W. (2021). Job screening service halts facial analysis of applicants. WIRED. https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/; Raji & Buolamwini, 2019; Raji, Gebru, et al., 2020.

CHAPTER 26

MITIGATING ALGORITHMIC BIASES THROUGH INCENTIVE-BASED RATING SYSTEMS

NICOL TURNER LEE

Introduction

Researchers and policymakers are positioning various resources to validate the trustworthiness and ethical design of artificial intelligence (AI) systems, including audits, assessments, badges, and other technical strategies. In the United States, the National Institute of Standards and Technology (NIST) has instituted best practices and some standardization around artificial intelligence (AI) systems and governance (Exec. Order No. 13859, 2019). However, a formal rating structure incentivizing the private and public sectors to advance more responsible AI has yet to exist, especially when it comes to identifying and mitigating offline and online biases.

Increasingly, AI systems are impeding the civil and human rights of subjects, largely due to the lack of protocols for the identification and mitigation of online biases. Mathematician Cathy O'Neil captured the societal implications of big data in her book, *Weapons of Math Destruction* (2016), which raised alarms on the unintended consequences of big data, artificial intelligence (AI), and other autonomous systems that rely on machines for predictive decisions. Author Safiya Noble (2018) examined how search engines perpetuate and reinforce hegemonic views of women and people of color. Through experimentation, Noble discovered that a search on the keywords "black girls" returned objectifying and pornographic results. Meanwhile, image queries on Google returned pictures of blonde, White women when searching the keyword "beautiful" and images of White males showed up for the keywords "professor style" (Noble, 2018, pp. 19–23).

While technologists have sought to address these forms of unconscious and explicit biases through model revisions like the replacement or omission of certain features and attributes, these biases continue to downgrade public confidence in automated decision-making tools.

That is why developing more diverse work teams and incentivizing those who create, license, or distribute algorithms to consider the unintended consequences of their models are critical in addressing online biases. It is equally important for computer and data scientists to understand that their values, norms, and assumptions seep into the design and execution of machine-learning (ML) models and can often go unchecked throughout the life cycle of a single or multiple algorithms. For example, an innocuous Google search of the words "happy teenagers" by Black students resulted in photos of happy, smiling White people, while Black teenage faces were delivered in mug shots (Guarino, 2016). While Google quickly remedied the problem in their search queries, other similar occurrences have occurred with companies using some of the same design protocols. Most recently, in September 2021, social media giant Facebook apologized after the company's AI labeled a video featuring a Black man as "primates" (Lyons, 2021).

In this chapter, I present a particular strategy for proactively addressing and mitigating online biases—incentive-based rating systems. Various applications, from ride- to home-sharing apps, have come to depend on them to evaluate the effectiveness of their services and increase trust among those accessing their platforms. E-commerce companies have also used digital rating systems, along with feedback ranking tools, to gather user input and amplify the more favorable comments from consumers. Despite being highly prevalent in the online, commercial marketplace, rating systems are not often incorporated into technical and policy proposals to make AI more equitable, fair, and inclusive. Certainly, the application of such rating systems will look different among algorithmic deployments. Yet, if computer and data scientists were interrogated by non-technical experts, or the people who are the subjects of their models, what significant progress could be made to thwart the marginalization of certain populations?

I also argue that what's missing in the suite of available resources to tackle online biases are more precise frameworks for questioning computer models around differential treatment, and potentially disparate impact on vulnerable users, especially when done by technical and non-technical experts. Increasingly, we have seen public calls for improved technical cadence among developers to retract online biases and create programmatic and proscriptive policy guardrails that protect consumers and offer more transparency of algorithmic impacts. A mechanism like a formal rating system can reward developers and companies that attempt to tackle these disparities by creating more collaborative and diverse work teams, as well as socio-technical best practices, that fulfil the public interest and social welfare goals of policymakers and consumers.

One consumer-facing framework that can be helpful to this discussion of rating algorithmic models is developed and led by the United States Environmental Protection Agency's (EPA), whose ENERGY STAR® Rating program (n.d.) converges governance, standards, and user expectations around the performance of certain consumer products, mainly appliances. The program is a partnership between the EPA, Department of Energy (DOE), Federal Trade Commission (FTC), and thousands of commercial, utility, state, and local organizations, to help consumers make more environmentally friendly decisions about a range of consumer appliances. Its goals are to imbue product trust and confidence, which are also worthy concerns for algorithms that determine one's eligibility for credit, housing, employment, and arrest. While these particular cases may also require some level of regulatory or legislative enforcement, the Energy Star Rating program brings clarity to the operational cadence and performance outputs, which are driven by agreed-upon industry

standards, policies, product disclosures, and consumer input. In an age where consumers are losing agency and control over their online behaviors, an incentive-based rating system or participatory framework could drive public accountability, and embolden what policymakers have recently articulated as a "rights-based approach to fairness" (Lander & Nelson, 2021).

More important, a proposed AI rating system not only rewards developers and companies that are diligent in their management and integration of fairness principles and practices in the model's design, but also ensures the robust interrogation of models that can have intended and unintended consequences on users, especially for members of historically disadvantaged groups. Examples of this might include disproportionate loan rejection, higher bail, or longer criminal sentences, among other sensitive use cases. Compared to the predominant focus by computer and data scientists on the technical mining of errors, this chapter argues for more non-technical or sociological inquiries of the operation and performance of ML algorithms that examine the societal impacts, the diversity of the team and data, and the real-world contexts in which models are deployed. In other words, a rating system that considers the efficacy and performance of proposed models by understanding: (1) who's involved in the developing and monitoring of the algorithm; (2) the representativeness and quality of the dataset used to train the algorithm; (3) the basis in evidence within the algorithm's design (4) the privacy protections afforded to consumers over the data collection process; (5) the usage of audits and disparate impact reviews to assess the algorithm's outcomes; and (6) the integration of consumer feedback loops throughout the algorithm's life cycle.

The next section explores such sociological construction for understanding online biases in ML algorithms before unveiling a series of case studies that surface the differential treatment of vulnerable populations, especially in the areas of employment, credit scoring, child welfare, public benefits, and policing. The chapter concludes with a series of questions that should be part of the design, execution, and evaluation processes of ML tools, which, if disclosed through a rating system, can heighten a user's trust not only in the code driving the model but also the process. In the end, a proposed rating system not only differentiates developers who exercise a better duty of care over algorithms guiding sensitive use cases, but also incentivizes the marketplace for improved algorithmic performance with a reduced risk of discrimination or other predatory outcomes.

Bias in AI-enabled Decision-making

In the 1950s, Arthur Samuel engineered the fundamentals of self-learning computers, whose sophistication continues to proliferate into emerging technologies. Contemporary algorithms, which are sets of step-by-step instructions that computers follow to perform a task, are integral to automated decision making (DeAngelis, 2014; Garbade, 2021). Yet the evolution in the complexity of modern-day algorithms have also made them more opaque. Many models not only account for the initial input variables defining users or other scenarios, but also take a large amount of data from other sub-features like experiences, patterns, and other historical variables or inferences. That is why algorithms are commonly known as "black boxes" because these "systems . . . hide their internal logic to the user" and

sometimes even its designers (Guidotti et al., 2018). When bias happens, it is often complex to discover or unravel its sources. Yet for consumers, these ML models continue to be critical in both essential and non-essential activities—whether for determining the positioning of movies on video streaming services or for gauging the effects of certain medications on blood pressure via cardio meter algorithms. In these and other cases, algorithms make decisions and predictions around human behaviors primarily based on existing and future data trails, primarily driven by the online surveillance of the personal and collective experiences of users and subjects.

Generally, such monitoring is innocuous. For example, online users who have extensive purchasing histories of "blue dresses" benefit from targeted online ads deliveries, videos, and other channels (e.g., social media groups) marketing these products to them. But market surveillance can quickly transition from seductive to harmful for consumers when the data associated with "blue dresses" combine with online proxies like zip codes, photos, and names, among other things, that predict behaviors beyond purchasing, or the algorithm's original intent. Early research from Harvard professor Latanya Sweeney, whose work evaluates the societal implications of technology, discovered online biases in various types of ads, finding that searches of "Black-sounding names," like DeShawn, Darnell, or Jermaine, were more likely to prompt users to search for arrest records (Sweeney, 2013, pp. 44–54). "White-sounding names," including Jill, Geoffrey, and Emma, elicited an opposite, neutral response.

That is why flushing racial discrimination from computer models helps to manage the seen and unforeseen reputational risks on the part of technology developers and the companies that license, sell, and distribute algorithms (Lange & Duarte, 2017). United States policymakers have intervened by drafting a plethora of congressional bills and state-led legal inquiries, and in some instances, followed the lead of European Union regulators who have begun to classify high-risk ML models. Civic-driven campaigns in the United States, like the recent efforts of "#StopHateForProfit" led by the NAACP and other social justice organizations, have also attempted to raise more public awareness (U.S. House Committee on Financial Services, 2021; U.S. Senator Cory Booker of New Jersey, 2019; Hern, 2020). But not all model fallibilities are driven by malfeasance on the part of computer developers. Even the slightest suggestion of differential treatment or disparate impact resulting from AI models contributes to their untrustworthiness by consumers and heightens the perceived lack of ethics emanating from them. In some cases, these errors also violate the civil and human rights protections afforded to historically marginalized groups, for example, criminal justice algorithms that increase the likelihood of higher bail limits imposed on Black defendants over White ones (Arnold et al., 2018).

These reasons, in part, make it more difficult to clearly identify and remedy online biases, especially in cases where the inferences cloud the ability to properly interrogate the ML. That is why without the appropriate combination of policy guardrails, transparency, and consumer feedback, the virtuous cycles of discrimination will not dissipate but worsen.

Defining machine learning bias

Turner Lee et al. (2021) define online bias as scenarios in which similarly situated people and objects are treated differently within models, despite the programmer's expectations

around the result. While some computer scientists have presented valid arguments around the technical sanitation of data inputs through more objective and "blind" modeling that reduces or averts online discrimination entirely, others have acknowledged that remedying long-standing, systemic inequalities via computer models is virtually impossible (Uzzi, 2020; Smith, 2019). Without positioning equity in the design phase of the AI, it is also difficult to mitigate such risks in its execution.

A comprehensive civil rights framework introduced by lawmakers and adopted by big tech companies could help to mitigate the constant misunderstandings around anti-discrimination laws, while establishing appropriate guidelines for fair and transparent online conduct that set standards for the equitable treatment of federally protected groups. But these applications would have to be applied to models that pose higher risks to consumers, which are the cases described below with regards to hiring, credit scoring, child welfare, public benefits, and policing, to name a few.

Emotion AI in hiring

Organizations are increasingly using "Emotion AI"—algorithms designed to read facial expressions, body language, voice patterns, eye movement, and more—to flag job candidates that exhibit specific qualities like determination, confidence, and empathy (Purdy et al., 2019). As a hiring tool, Emotion AI can limit opportunities for advancement among historically disadvantaged communities because emotions are not easily measurable and may vary between people of different ages and races (Fölster et al., 2014, p. 30; Rhue, 2019). Professor Lauren Rhue (2018) conducted a study using 400 images of NBA players and found that two popular emotion recognition products consistently assigned Black players more negative emotions than White players. Despite these flaws, HireVue, a company specializing in Emotion AI, continues to sell the product to major employers like Goldman Sachs, Intel, Amazon, Target, and JP Morgan to evaluate job candidates (O'Brien, 2021). In 2021, the company removed facial expression analysis from its hiring algorithm after the Electronic Privacy Information Center lodged a complaint with the FTC, arguing that the company's use of Emotion AI in its hiring software amounted to an unfair and deceptive trade practice (Harwell, 2019). But despite these and other complaints, Emotion AI software continues to be widely deployed and largely unregulated in the human resources industry.

Loan and credit-scoring software

Contemporary algorithms have largely replaced human loan officers in evaluating creditworthiness (Miller, 2020). In some ways, FinTech algorithms have reduced discrimination in loan approval and pricing, making credit more widely available to historically disadvantaged populations, exhibiting 40 percent less price discrimination than face-to-face lenders and no discrimination on loan approvals (Bartlett et al., 2019). However, low-income and disadvantaged populations are still less likely to generate the types of data that are used in credit-scoring because they have less internet connectivity and less capital to spend. In fact, 44 million Americans are "'credit invisible' because they are "disconnected from mainstream financial services and thus do not have a credit history" (Gilman, 2020, p. 11).

When FinTech algorithms fail to take that invisibility into account, they perpetuate historical patterns of discrimination in the financial services sector.

Child welfare determinations

As of 2018, child protective authorities in at least twelve states have developed or deployed algorithms for predicting the likelihood of child abuse (Saxena et al., 2020). Algorithmic bias presents tangible threats related to fairness within the child welfare system. Many of the variables analyzed—such as county jail bookings and interactions with the juvenile justice system—are produced by institutions that disproportionately target people of color. From a child welfare perspective, this generates two parallel dangers: not only will Black and Hispanic children be unnecessarily investigated and removed from nurturing homes, but White and wealthy children may not be removed from their homes despite more dangerous family environments (Saxena et al., 2020). Because algorithms are considered "black boxes," there is also a lack of transparency in how crucial child welfare decisions are made or predicted. Even when algorithmically generated risk scores are one of several factors in a child welfare determination, judges and case workers often accord algorithms greater weight in their decision-making due to their perception of algorithms being objective measures. A meta-analysis of 50 peer-revied publications on child welfare algorithms revealed that most algorithmic models were not designed based on child welfare literature and used variables with little or "no predictive validity" (Saxena et al., 2020, p. 9). Consequently, these models that are making life-changing decisions for children and their families are used without transparency, evidence, or, in some instances, objective data to generate reliable risk factors.

Eligibility for public benefits

Several states are using algorithms to identify fraudsters and ineligible recipients to eliminate fraud, waste, and abuse in social safety net programs. In 2013, Michigan introduced the Michigan Integrated Data Automated System (MiDAS) to identify fraud in unemployment insurance benefits. The system falsely accused approximately 48,000 unemployment insurance recipients of fraud and required them to pay back the benefits with interest and a civil penalty of four times the amount they had received (Gilman, 2020). Similarly, Indiana entered a $1.3B contract with IBM to algorithmically disburse benefits for social safety net programs. Over the course of three years, the algorithm denied benefits to one million people, representing a 54 percent increase over the previous three years (Gilman, 2020, p. 38). The onset of the COVID-19 pandemic ushered in an even swifter transition to automated fraud detection. As state unemployment agencies were flooded with claims at the peak of the pandemic in March 2020, 27 states entered contracts with the private sector firm, ID.me, to provide identity authentication through its facial verification software (Metz, 2021). However, the use of this software has proven controversial with one of its customers, the Internal Revenue Service discontinuing its use for tax returns and processing (Picchi & Ivanova, 2022). The other part of the problem in government's use of facial detection and recognition systems is the technical inaccuracies in the identification of people of color. For example, the results of women and people of color using ID.me for

unemployment verification were more likely to result in denials or rejections in the state of Florida, who has consistently relied on these systems for identity verifications (McGivern, 2021). However, the United States is not the only country that has begun automating safety net programs with disastrous results for the poor. Automated public benefits systems in the United Kingdom, Australia, the Netherlands, and Denmark are also adversely discriminating against and disempowering vulnerable communities (Marsh, 2020; Terzis, 2017; Richardson et al., 2019; Mchangama & Liu, 2018).

Facial recognition systems and policing

In 2016, the Georgetown Law Center on Privacy and Technology found that law enforcement agencies across the United States have access to facial image databases encompassing over 117 million Americans, or over one-half of all American adults. They also concluded that one-quarter of all local and state police departments had the ability to run facial recognition searches despite facial recognition software demonstrating clear algorithmic bias (Garvie et al., 2016). The *New York Times* has identified three instances in which facial recognition technology have led to the wrongful arrests of Black men—although the real number is likely much higher because some states do not require law enforcement to disclose when facial recognition technology is used to identify a suspect (Hill, 2020; Valentino-DeVries, 2020). Wrongful arrests driven by facial recognition technologies illustrate that without strong regulation, the technology will continue to reinforce and exacerbate racial profiling and biased policing. They also surface the challenges that vulnerable populations face when discerning biases primarily found in online platforms, products, and services, especially when used by law enforcement.

What the case studies reveal

Systemic barriers to equal opportunities predate the use of algorithmic models and enable discrimination. But the case studies suggest the possibility of even greater precision in discriminatory and predatory online behaviors with recent developments. Further, companies like Facebook and Uber have surfaced the adverse effects of more blatant and explicit discrimination, especially in the areas of housing options and accommodations. In 2018, the United States Department of Housing and Urban Development (HUD) agency filed a complaint against Facebook for allowing home sellers and landlords to pick and choose who could see their ads, which was in violation of the Fair Housing Act that forbids targeted advertising based on federally protected traits (Zhou, 2018). According to a 2016 study by Stanford University, MIT, and the University of Washington, Uber drivers canceled rides twice as often for men with Black-sounding names (Newcomer, 2016), which mirrored the historic denial of services to Blacks from slavery to today. In both cases, it is a fair assumption that some evidence of disparate impact was at play, only this time, it played out as a form of digital discrimination, where online resources posed barriers to equitable access.

Both companies responded to such grievances with self-regulatory actions to lessen the public volatility around the complaints. But they did not directly solve the problem. In April 2021, a study found once again that Facebook's ad algorithms continued to exclude

women from certain job opportunities (Hao, 2021). While Uber recalibrated their algorithm to allow for more blind selection of drivers and riders, recent studies found that discrimination persisted against LGBTQ+ and nonwhite people (Srikanth, 2020). Once again, the models contributed to more precise targeting and adverse surveillance of certain populations, calling into question the sanctity of the civil rights regime created to protect them from such harms.

Profit Over Performance?

Some researchers and civil society organizations have argued that profit incentives in the tech marketplace often facilitate unfair outcomes for vulnerable populations, making self-regulatory actions without enforcement insufficient to find and mitigate online biases. But regulation or other actions by the government may not sufficiently change or counter bad actors in certain marketplaces (Winston, 2008). That is why, in part, an incentive-based rating system may be able to effectuate some timely changes to high-risk scenarios where AI systems are at work.

The EU's Artificial Intelligence Act (2021) uses the term "high-risk AI" to refer to systems that pose "risks to . . . health or safety or to the protection of fundamental rights." The notion of "high-risk AI" is one that should also be adopted by the United States in its exploration of fairness and ethics of systems used in sensitive use cases. Further, a rating system may need legible enforcement, disclosures, or consumer appeals to redress implicit and explicit attacks on user livelihoods, including user retribution on automated decisions or information channels that reveal the use of AI in predictive decisions. My point is that any rating system that is entirely voluntary may not comport with the ultimate goals of addressing and mitigating online biases. Rather, balancing the tensions between a developer or industry's desires for profit over performance, or vice versa, may require some levels of enforcement via appropriate policy guardrails, consumer input, and more frequent and reasonable independent or third-party audits of algorithms and other AI systems. All these factors comprise the elements of a proposed rating system.

How Ratings Systems Have Mitigated Risk in Other Industries

Rating systems initiated by government actors are not new in the United States, although they tend to be more static and established around the validation of either models or standards. Since 2019, the Centers for Medicare and Medicaid have implemented a Five-Star Quality Rating system that helps consumers, their families, and caregivers select and compare nursing homes (Centers for Medicare & Medicaid Services, 2019). The United States Food and Drug Administration generates nutrition labels that validates the stated nutritional values of certain food items for consumers to make informed choices about their consumption (U.S. Food and Drug Administration, 2020).

The Energy Star Rating program

In 1992, the EPA started a voluntary federal program as part of the Clean Air Act. Coined the Energy Star, the program was designed to save money, reduce environmental erosion, and improve energy efficiency (ENERGY STAR®, n.d.). The program identifies and promotes energy-efficient products (e.g., clothes washers, refrigerators, and other specific appliances) and buildings that protect the environment. It also advises consumers through Energy Guide labels, which are the yellow tags attached to most home appliances. The label helps consumers identify high-efficiency models and find companies working to meet EPA standards. The FTC works in partnership with the EPA and the DOE in the implementation of the Energy Star program and serves as the enforcer for companies that fail to comply with the guidelines. In 1979, the FTC adopted the Appliance Labeling Rule, which requires manufacturers to disclose energy information about major household appliances to help consumers compare the energy use and efficiencies of competing models (Federal Trade Commission, 2012). Overall, the program is collaborative and participatory, with thousands of entities from product manufacturers, national, regional, and local retailers, utilities, state energy offices, industry trade associations, and financial companies working with the federal government on the validation of processes, standard, and outputs.

What is more relevant about this government program is the targeted focus on increased consumer awareness when purchasing highly efficient appliances. The visible EnergyStar logo and yellow label help consumers distinguish between the good and bad actors in the industry, the latter of which make products that would otherwise degrade the environment in the long-term and badger consumers with unexpected and exorbitant energy expenses. Nearly one-fourth of all energy consumed in the United States can be attributed to households, amounting to $200 billion per year spent by consumers (U.S. Government Accountability Office, 2007). The federal government realized the increase in costs could be largely contained through standards and a set of best practices that incentivize Energy Star program participants to innovate for the public good, reducing consumer costs and environmental impacts.

Rating Systems for Algorithms

An analogous energy star rating for algorithms could replicate what the EPA has done and involve various actors in the design of transparent standards or development of universally accepted codes of conduct that consumers should come to expect when using ML algorithms. Figuratively, the energy-star rating for algorithms could evolve in concept from NIST, the FTC, or similar regulatory agencies around the world to: (1) assess and evaluate the ethical consequences of algorithmic models *before* they are deployed in settings where they have a real impact on human lives; (2) bring more transparency and disclosure to the purpose, design, and outcomes of certain algorithms; (3) improve accountability mechanisms from the developers and companies that license, distribute and deploy them, including impact assessments, audits; and (4) create a channel for robust consumer feedback loops between developers and users to continuously verify the model's performance.

More important, the framework can avert overt discrimination that may appear implicitly and explicitly in ML models.

As a participatory framework that motivates the balance between industry self-regulation, policy guardrails, consumer feedback, and some type of enforcement or retribution, the proposed rating system for AI is a legible provocation to incentivize technology developers to work more inclusively and efficiently to quell online biases, while expanding the rubric of participation for others who are equally vested in the model's efficacy, including social scientists, privacy experts, civil society organizations, and consumers. It restores some level of consumer control, infuses achievable standards, and improves upon technical efficacies that optimize for diversity and not simply for discrete computer models.

Such external recognition for fair algorithms can appear as a reputational badge similar to the one launched by the Equal AI consortium and the World Economic Forum (EqualAI®, n.d.). Alternatively, a public statement, or written disclosure, on the process could be made available to users and government agencies. Whatever the mechanism to produce more equitable AI, consumers and policymakers must be able to easily know that developers and companies have worked to address concerns of online biases.

As explained by Fung et al. (2007), targeted transparency has historically been proven to increase incentives for vendors to do better, while ensuring that consumers also have access to the quality services they deserve. When applied to AI, rating systems can facilitate better choices for consumers, helping them steer clear of lazily developed and tested algorithms that, in turn, thwart their livelihoods through the violations of human or civil rights. Such thinking could be applied to the credit-scoring or public benefits eligibility algorithms previously mentioned. Increased transparency and disclosure of the practices in the model's design can also pressure companies to do better, knowing that a badly graded algorithm could directly impact how appealing they are to customers, and thereby lower their profit margin or trustworthiness. While comparing the EPA's model with that of AI may seem unintuitive, the intent is the same, which is to provide more agency for developers, the government, and consumers to develop models optimized for public good, and not just for profits and performance.

Components of an AI-applied rating system

When constructing a rating system for algorithms, it is important to note that it is not only addressing the technical concerns related to computer models, but also the sociological implications of automated or predictive decisions. My research does not seek to challenge the technical cadence of developers, especially mathematicians, computer, and data scientists, who work diligently to create objective ML models. Instead, the proposed rating structure invokes a deeper look at some of the non-technical conditions that are in place in the pre-planning of AI systems, the initial launch and execution, and evaluation throughout the algorithm's lifecycle. Harvard researcher Jinyan Zang has coined his process, "fairness through awareness," where a full inventory of existing and potential threats is analyzed by a robust team of designers and decision makers that know the importance of context in the deployment of automated decision-making tools (Zang, 2021).

In this vein, the proposed framework, which could either be a voluntary best practice or enforceable by regulatory agencies, presents questions that developers, companies,

government, and civil society should be asking as they consider checks and balances in the algorithmic economy.

1. *What is the diversity of the workforce team developing the algorithm?* Having representative voices when planning and launching algorithms will be instrumental in helping discover data insufficiencies, unintended consequences, and potential areas leading to disparate impact based on race, gender, sexual orientation, or disability. In addition to having individuals of diverse demographic backgrounds, other interdisciplinary assets, including social scientists, legal experts, and citizens, can either offer or relate to the "lived experiences" of subjects while contextualizing the model in a rather opaque digital sphere. According to the AI Now Institute, companies like Facebook and Google have a nominal representation of women in their research staff; 15 percent and 10 percent, respectively (West et al., 2019, p. 3). Among Black workers in big tech companies, the numbers are more alarming, with only 2.5 percent of Google's workforce classified as Black, and four percent at Facebook and Microsoft (West et al., 2019, p. 5). Problems are not limited to the talent pipeline, as "[w]orkers in tech companies experience deeper issues with workplace cultures, power asymmetries, harassment, exclusionary hiring practices, unfair compensation, and tokenization that are causing them to leave or avoid working in the AI sector altogether" (West et al., 2019, p. 3). A rating approach could reward design teams that value diversity and inclusion, especially experts from under-represented backgrounds or those most affected by the ML algorithm. Drawing more experts into the conversation, including those from non-technical backgrounds like the social sciences, can encourage more balanced and expansive interpretations of AI models and their outcomes.
2. *Does the algorithm's design have a basis in evidence?* The fallibilities in child welfare determinations present reasons for exploring the basis in evidence of algorithmic models. In the citation to the meta-analysis of 50 peer-reviewed studies of child welfare algorithms, researchers found that most of the predictors used in child welfare algorithms were "easily available and readily quantifiable" but many "had not been properly validated and [could] lead to unreliable predictions" (Saxena et al., 2020, p. 8). When assessing the various scenarios of an algorithm's decision, developers should have some references rooted in social science or related scientific literature that creates an aperture for greater understanding of the research problem and execution. By grounding their design in scientific inquiries or social science research, attuned developers will enhance their perspectives before engaging in the model's design, or, at minimum, they could enlist an interdisciplinary team of researchers and other partners to encourage deeper understandings. Herein, rating systems should consider whether there is a strong theoretical basis for the algorithm's design, and whether the developer exercised their diligence or gathered multidisciplinary research to explore the subject at hand and anticipate potential blind spots.
3. *Is the algorithm trained on representative and quality datasets?* Some researchers argue that the primary problem of racial biases and other systemic inequalities rests with the data used to train the algorithm. While data training AI systems are often right sized for the problem being solved, certain types of training data can be wrong, non-representative, and outright inappropriate. In other words, training data and quality are only as good as the sources from which they are derived. In March 2020, a prominent

group of researchers released a paper titled "Datasheets for Datasets," in which they proposed a standard set of questions that must be answered for every widely used dataset to better flag datasets that may not be inclusive (Gebru et al., 2020). The article spoke to the lack of standardization and consistency among training datasets to explicate the information's operating characteristics, test results, recommended uses, and other useful inputs. Every dataset should be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Added to these assumptions should also be disclosures about the data inadequacies or shortcomings, like those associated with the face detection of people of color and women (Buolamwini & Gebru, 2018; Garvie et al., 2016). Without fully vetting equity in the proposed data, there will be difficulty in mitigating such risks at the onset of the ML algorithm's development and throughout the life cycle, especially when masked by online proxies, such as zip codes and photos, and other factors that can potentially facilitate discrimination. That is why data transparency statements, perhaps through written disclosures that amplify sources of weakness, could surface where the challenges arise in ML models.

4. *Does the data collection process respect privacy?* Algorithms used to make life-altering decisions in the public sector often draw on data from unexpected sources, such as mental health records and eligibility for public benefits (Glaberson, 2019). The use of these data sources may feel like a betrayal of privacy for those who are impacted by the model's decision-making capacities. However, information fed into computational models must preserve and protect individuals' privacy in both inputs and outputs. The United States currently lacks federal privacy protections, instead relying on a patchwork of state-enforced laws. Following the lead of other countries like the EU, the United States must adopt legislation that lays out the specifics for how data can be collected, used, shared, and stored. Technology companies and other developers who clearly disclose their privacy-enhancing actions could experience higher ratings under an incentive-based framework, and potentially differentiate themselves in the marketplace from others with ill-regard.
5. *Does the company/organization conduct audits and disparate impact reviews?* Audits and impact assessments, especially those geared towards the determination of either differential treatment or disparate impact of the ML algorithm, add to positive ratings when used to validate the measurements built into a model's design and outputs. Brookings scholar Alex Engler offered a helpful set of criteria for algorithmic audits, through including auditor or third-party independence, representative analysis, and potential adversarial actions like nonapparent data manipulation (Engler, 2021). He also argued that scientists may desire access to modeling dependencies and documentation, as well as the ability to deploy various use cases under secondary, primary, and even tertiary datasets and conditions to ensure the evenness of outcomes. Dependency analysis can assist in the scrutiny of pretrained ML models for inherent biases, and documentation—as mentioned previously—can help locate the errors. More important, certain algorithms, including those used for hiring or criminal justice decisions, should be regularly reviewed around its outcomes, considering their impacts on more vulnerable populations, data qualities, fairness tradeoffs and short- and long-term results. Audits and impact assessments can also be useful in determining whether proxies have been enabled at some point in the model's execution,

or if there exist any other adverse impacts on federally protected groups and their characteristics.

6. *How receptive is the company to consumer feedback and further model revisions?* The engagement of consumers in rating systems can be a useful resource in gauging trust and confidence in particular algorithms. In AI, consumers are essentially the products, or are providing the attributes or features for models. Consumers are not always involved in providing feedback and suggestions to developers and industries licensing and distributing algorithms. But emerging technologies and datasets cannot be optimized for diverse populations if their voices are not included, and, in some case, intentionally excluded, making the case for reduced discriminatory outcomes. Turner Lee et al. (2021) has argued that policymakers permit regulatory sandboxes as part of improved sanitation of potentially biased AI models. Regulatory sandboxes can invite consumers into the model's development, allowing for innovation both in technology and its regulation because they involve some level of partnership between government, private industry, or academics. But the primary reason for establishing some type of consumer feedback loop is to return human agency to the evaluation and validation of ML models back to them, while also conforming to the premise that users have fundamental roles and rights in the digital economy. Embracing civil society and consumer engagement in AI development starts in undergraduate education where computer and data scientists should learn about inclusive methodologies that promote ethics, equity, and fairness. Students in the data and computer sciences should also design to compliance, establishing best practices around civil rights, especially in their application to banking, health care, education, and other sensitive use cases that closely correlate with positive outcomes for marginalized groups. As these ML models mature, having descriptive guardrails and, in some instances, proscriptive regulatory and legislative guidance for developers and companies around behavioral, associational, and/or inferential models can also help limit, and potentially repudiate, online discrimination, which can also be re-directed in the collection of user experiences with online models.

Taken together, these six questions, which are not exhaustive, contribute to the interrogation of algorithms and provide some basis for increased reputational trust by demonstrating some diligence in vetting and identifying vulnerabilities for ML algorithmic biases. Like the Energy Star Rating program, the framework rests upon some work being done by technology developers from the beginning and throughout the model's lifecycle—whether in the form of universal technical and sociological cadence, or by way of more rigorous inquiries.

Conclusion

For too long, unregulated algorithmic development has perpetuated and exacerbated biases against historically marginalized populations, adversely impacting their lives in ways that they have no knowledge of nor control over. A rating system approach, like the EPA's Energy Star Rating program, can be an important first step in predicting and solving online biases.

But this will only come to be if developers and companies are incentivized to prioritize equity and inclusion in the interest of managing their online reputational risks. Modern day algorithms are also at a point where consumers need to be granted the much-needed agency and transparency to understand and correct the algorithms that determine critical aspects of their lives. With the ever-growing presence of technology and the dominance of automated systems in our daily lives, the sociological implications of AI models and the elevated concerns for the welfare of consumers must be considered to drive responsible innovation.

Acknowledgments

Special acknowledgments in support of this research chapter go to Samantha Lai and Emily Skahill. The opinions in this chapter are solely from the author and do not represent the thoughts, ideas, or opinions of the Brookings Institution.

References

Arnold, D., Dobbie, W., & Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics 133*(4), 1885–1932.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-lending discrimination in the fintech era* (No. w25943). National Bureau of Economic Research. https://doi.org/10.3386/w25943.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research, 81:77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Centers for Medicare & Medicaid Services. (2019). *Five-Star Quality rating system*. U.S. Department of Health and Human Services, October 7. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandComplianc/FSQRS.

DeAngelis, S. F. (2014). Artificial intelligence: How algorithms make systems smart. *Wired*, September 5. https://www.wired.com/insights/2014/09/artificial-intelligence-algorithms-2/

ENERGY STAR®. (n.d.). *Energy Guide vs. ENERGY STAR*. U.S. Environmental Protection Agency and U.S. Department of Energy. https://www.energystar.gov/products/ask-the-experts/whats-the-difference-between-the-energy-guide-and-energy-star.

Engler, A. (2021). *Auditing employment algorithms for discrimination*. The Brookings Institution. https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/.

Exec. Order No. 13859, 84 Fed. Reg. 3967 (February 11, 2019).

Fölster, M., Hess, U., & Werheid, K. (2014). Facial age affects emotional expression decoding. *Frontiers in Psychology 5*. https://doi.org/10.3389/fpsyg.2014.00030.

FTC Amends Appliance Labeling Rule to Ease Burdens on Businesses. (2012). Federal Trade Commission, December 31. https://www.ftc.gov/news-events/press-releases/2012/12/ftc-amends-appliance-labeling-rule-ease-burdens-businesses.

Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge University Press.

Garbade, M. J. (2021). Clearing the confusion: AI vs machine learning vs deep learning differences. *Medium*, April 18. https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb.

Garvie, C., Bedoya, A., & Frankle, J. (2016). Perpetual line up. Georgetown Law Center on Privacy and Technology, October 18. https://www.perpetuallineup.org/background.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for datasets. *Communications of the ACM 64*(12), 86–92.

Gilman, M. (2020). *Poverty lawgorithms: A poverty lawyer's guide to fighting automated decision-making harms on low-income communities*. Data & Society. https://datasociety.net/library/poverty-lawgorithms/.

Glaberson, S. K. (2019). Coding over the cracks: Predictive analytics and child protection. *Fordham Urban Law Journal 46*(2), 307.

Guarino, B. (2016). Google faulted for racial bias in image search results for black teenagers. *Washington Post*, June 10. https://www.washingtonpost.com/news/morning-mix/wp/2016/06/10/google-faulted-for-racial-bias-in-image-search-results-for-black-teenagers/.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys 51*(5), Article 93. https://doi.org/10.1145/3236009.

Hao, K. (2021). Facebook's ad algorithms are still excluding women from seeing jobs. *MIT Technology Review*, April 9. https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/.

Harwell, D. (2019). Rights group files federal complaint against AI-hiring firm HireVue, citing "unfair and deceptive" practices. *The Washington Post*, November 6. https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/.

Hern, A. (2020). Facebook to be hit by its largest ever advertiser boycott over racism. *The Guardian*, June 24. http://www.theguardian.com/business/2020/jun/24/ben-and-jerrys-joins-facebook-advertising-boycott-racism.

Hill, K. (2020). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*, December 29. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

Lander, E., & Nelson, A. (2021). Americans need a bill of rights for an AI-powered world. *Wired*, October 8. https://www.wired.comstory/opinion-bill-of-rights-artificial-intelligence/.

Lange, A. R., & Duarte, N. (2017). Understanding bias in algorithmic design. *Medium*, September 6. https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e.

Lee, N. T., Zang, J., & Harrison, D. (2021). Online behavioral advertising can be discriminatory, so what's the plan to fix this? Online podcast episode, October 4. https://www.brookings.edu/blog/techtank/2021/10/04/techtank-podcast-episode-29-online-behavioral-advertising-can-be-discriminatory-so-whats-the-plan-to-fix-this/.

Lyons, K. (2021). Facebook says its AI mislabeling a video of Black men as "primates" was "unacceptable." *The Verge*, September 4. https://www.theverge.com/2021/9/4/22657026/facebook-mislabeling-video-black-men-primates-algorithm.

Marsh, S. (2020). Councils scrapping use of algorithms in benefit and welfare decisions. *The Guardian*, August 24. http://www.theguardian.com/society/2020/aug/24/councils-scrapping-algorithms-benefit-welfare-decisions-concerns-bias.

McGivern, K. (2021). Facial recognition meant to stop unemployment fraud is blocking legitimate applicants. *ABC Action News WFTS Tampa Bay*, June 7. https://www.abcactionnews.com/news/local-news/i-team-investigates/facial-recognition-meant-to-stop-unemployment-fraud-is-blocking-legitimate-applicants.

Mchangama, J., & Liu, H. (2018). The welfare state is committing suicide by artificial intelligence. *Foreign Policy*, December 25. https://foreignpolicy.com/2018/12/25/the-welfare-state-is-committing-suicide-by-artificial-intelligence/.

Metz, R. (2021). Want your unemployment benefits? You may have to submit to facial recognition first. *CNN*, July 23. https://www.msn.com/en-us/news/us/half-of-us-states-are-now-using-facial-recognition-software-from-this-little-known-company-to-vet-unemployment-claims/ar-AAMtC1Y?ocid=msedgntp.

Miller, Jennifer. (2020). Is an algorithm less racist than a loan officer? *The New York Times*, September 18. https://www.nytimes.com/2020/09/18/business/digital-mortgages.html.

Newcomer, E. (2016). Study finds racial discrimination by Uber, Lyft drivers. *Chicago Tribune*, October 31. https://www.chicagotribune.com/business/blue-sky/ct-racial-discrimination-uber-lyft-study-blm-bsi-20161031-story.html.

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

O'Brien, M. (2021). Computerized job interviews: AI may judge you, determine whether you get hired. *Chicago Tribune*, June 15. https://www.chicagotribune.com/business/ct-biz-ai-online-job-interview-computer-20210615-iuukk7z3jzbqvbc7khk55em2qq-story.html.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Picchi, A., & Ivanova, I. (2022). ID.me says users can delete selfies following IRS backlash. *CBS*, February 9. https://www.cbsnews.com/news/irs-id-me-delete-facial-recognition-tax-returns-backlash/.

Purdy, M., Zealley, J., & Maseli, O. (2019). The risks of using AI to interpret human emotions. *Harvard Business Review*, November 18. https://hbr.org/2019/11/the-risks-of-using-ai-to-interpret-human-emotions.

Rhue, L. (2018). *Racial influence on automated perceptions of emotions*. *SSRN*, November 9. https://dx.doi.org/10.2139/ssrn.3281765.

Rhue, L. (2019). Emotion-reading tech fails the racial bias test. *The Conversation*, January 3. http://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test-108404.

Richardson, R., Schultz, J. M., & Southerland, V. M. (2019). Litigating algorithms 2019 US report: New challenges to government use of algorithmic decision systems. AI Now Institute. https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf.

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2020). A Human-Centered Review of the Algorithms used within the U.S. Child Welfare System. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376229.

Smith, C. S. (2019). Dealing with bias in artificial intelligence. *The New York Times*, November 19. https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html.

Srikanth, A. (2020). Despite changes, LGBTQ+ and racial discrimination persists in Uber, Lyft. *The Hill*, July 30. https://thehill.com/changing-america/respect/equality/509817-despite-changes-lgbtq-and-racial-discrimination-persists-in.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM 56*(5), 44–54. https://doi.org/10.1145/2447976.2447990.

Terzis, G. (2017). Austerity is an algorithm. *Logic Magazine*, *3*, 101–160 (December 1). https://logicmag.io/justice/austerity-is-an-algorithm/.

Turner Lee, N., Resnick, P., & Barton, G. (2019). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. The Brookings Institution. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

U.S. Food and Drug Administration (2020). The new nutrition facts label. U.S. Department of Health and Human Services, December 28. https://www.fda.gov/food/nutrition-education-resources-materials/new-nutrition-facts-label.

U.S. Government Accountability Office. (2007). *Energy efficiency: Opportunities exist for federal agencies to better inform household consumers,* September. (Publication No. GAO-07-1162). Washington, D.C.: U.S. Government Printing Office. https://www.gao.gov/products/gao-07-1162.

U.S. Government Accountability Office. (2020). *Facial recognition technology: Privacy and accuracy issues related to commercial uses*, July. (Publication No. GAO-20-522). Washington, D.C.: U.S. Government Printing Office. https://www.gao.gov/products/gao-20-522.

U.S. House Committee on Financial Services. (2021). *Waters and foster urge regulators to ensure algorithmic bias does not occur in emerging technology*. Press Release, November 29. https://financialservices.house.gov/news/documentsingle.aspx?DocumentID=408850

U.S. Senator Cory Booker of New Jersey. (2019). Booker, Wyden, Clarke introduce bill requiring companies to target bias in corporate algorithms. Press Release, April 10. https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms.

Uzzi, B. (2020). A simple tactic that could help reduce bias in AI. *Harvard Business Review*, November 4. https://hbr.org/2020/11/a-simple-tactic-that-could-help-reduce-bias-in-ai.

Valentino-DeVries, J. (2020, January 12). How the Police Use Facial Recognition, and Where It Falls Short. The New York Times. https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminatory systems: Gender, race, and power in AI*. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf.

Winston, C. (2008). The efficacy of information policy: A review of Archon Fung, Mary Graham, and David Weil's *Full Disclosure: The Perils and Promises of Transparency*. *Journal of Economic Literature 46*(3), 704–717. http:www.aeaweb.org/articles.php?doi=1-.1257/jel.46.3.704.

Zang, J. (2021). Case studies in public interest technology. PhD diss., Harvard. University Graduate School of Arts and Sciences. https://dash.harvard.edu/handle/1/37368422?show=full.

Zhou, Marrian. (2018). Facebook takes heat from HUD over allegedly discriminatory housing ads. *CNET*, August 17. https://www.cnet.com/news/facebook-takes-heat-from-hud-over-allegedly-discriminatory-housing-ads/.

CHAPTER 27

ROLE AND GOVERNANCE OF ARTIFICIAL INTELLIGENCE IN THE PUBLIC POLICY CYCLE

DAVID VALLE-CRUZ AND
RODRIGO SANDOVAL-ALMAZÁN

Introduction

Artificial intelligence (AI) has been adopted in different aspects and activities of our daily lives (Campion et al., 2020) and has the potential to enhance economic and social welfare, improving living conditions and giving an opportunity towards the understanding of complex problems (Cath, 2018). Governments can leverage their activities through devices and technologies that enable them to manage and control data, especially from citizens. Consequently, AI implementation incorporates characteristics different from traditional information technologies, such as bio-inspiration, autonomy, adaptive nature, and complexity. This situation implies the need to know their potential, as well as the dimensions in which they will have a direct impact on society, public policies, and governance (Criado et al., 2020; Wirtz et al., 2019). Now, AI has begun to be of interest to governments around the world, generating interest in analyzing the implication for governance and the public policy cycle. While AI continues evolving and its adoption in governments continues growing, its scope and how far it will go are still unknown (Zuiderwijk et al., 2021).

In this regard, AI's governance and the potential dynamic policy cycle is beginning to be of interest to the public sector agenda, especially in governments that have driven their policies with the support of the technology intensive use (Dafoe, 2018; Valle-Cruz et al., 2020; Wirtz et al., 2019). The use of AI by public organizations can facilitate or prevent unnecessary harm (Young, Himmelreich, et al., 2019), having the potential to further improve the public policy process providing benefits to governments, citizens, and the private sector (Janssen et al., 2020; Sun & Medaglia, 2019). Nevertheless, there is a potential danger that the intensive use of AI in government will result in a catastrophic dystopia that leads to the

dark side (Warren & Hillas, 2020; Wirtz et al., 2020). For this reason, it is essential to study the governance and the policy cycle on AI to mitigate its impacts in areas such as labor displacement, access inequality, oligopolistic market structure, reinforcement of totalitarianism, and strategic stability at the international level (Dafoe, 2018).

Public policies represent a set of objectives, decisions, and actions carried out by a government to solve social and economic problems. The solution to these kinds of problems can be supported by artificial intelligence algorithms and machines to improve decision-making and make government processes more efficient (Valle-Cruz et al., 2021). Public policies are also designed to solve wicked problems, as stated by Churchman (1967). More recently Paquet and Schertzer (2020) define the Complex Intergovernmental problem, like the one which involves several agencies to operate simultaneously to work for a common solution on a multi-causal problem. In this regard, a contemporary problem that has been a priority is the COVID-19 pandemic, which has generated several dilemmas throughout the world (Barbieri et al., 2021; Rodríguez-Rodríguez et al., 2021). Related to the pandemic, the public sector—in collaboration with the private sector—some NGOs, and civil society took on the task of providing different types of solutions to combat the crisis caused by the SARS-CoV-2 coronavirus, providing the opportunity to streamline the public policy cycle (Baxter & Casady, 2020; Park & Chung, 2021; Sarkar, 2021; Storr et al., 2021). It is here where different nations around the world made an emerging and improvised implementation of public policies and tools to combat the pandemic (Cantú et al., 2020; Coveri et al., 2020; Li et al., 2021); one of them is through the support of artificial intelligence (Capucha et al., 2020; Jr et al., 2020).

AI has the potential to improve the public policy cycle (Valle-Cruz et al., 2020). The inherent complexity of implementing AI in government has led the organizations' leaders to require the necessary knowledge to govern AI and to design public policies for the benefit of humanity. For this reason, it is urgent to analyze the AI governance: devising strategies, solutions, global rules, public policies, and institutions to ensure the AI benefits (Dafoe, 2018). Given the above, this chapter aims to analyze public policy in the age of AI. The chapter remainder consists of four sections. First, we will explore the public policy cycle. Then, we will describe the public policy cycle framework in the AI era. Finally, we will analyze the public policy cycle for COVID-19. We will end by providing conclusions and final comments.

Public Policy-Cycle Overview

A public policy represents a set of activities or rules emanating from one or more actors vested with public authority. A variety of governmental and non-governmental actors are involved in its development. Governmental actors may belong to one or more government levels. Non-governmental actors (such as trade unions, NGOs, associations, private sector, etc.) may also operate at different government levels. The public policy cycle is a set of decision stages and actions. It is a reference framework useful to make sense of the decision flow and procedures that make up a public policy (Knoepfel et al., 2003). The notion of the public policy cycle is an analytical device, intellectually constructed, for the purpose of

modeling, ordering, explaining, and prescribing a policy (Villanueva et al., 2007). Public policies represent a course of action and flow of information related to public goals, democratically defined, developed by the public sector, and frequently with the participation of the private sector (Lahera, 2004).

Public policies are specific solutions on how to manage public sphere issues. Although there are different frameworks related to the public policy cycle, most agree on four stages: agenda setting, policy formulation and decision making, policy implementation, and policy evaluation (Jann & Wegrich, 2017; Pencheva et al., 2020; Thierer et al., 2017). Each stage of the public policy cycle represents an analytical moment of heterogeneous quality and duration. However, the phases of the cycle are interdependent. Each of them is described below.

In *agenda setting*, problems and alternative solutions gain or lose the attention of the public and elites. Competition among groups to set the agenda is fierce because no society or political institution can address all possible alternatives to all problems (Birkland, 2017). Public problems are circumstances that have or will have adverse effects on society. To be considered as such, they must be widely spread and accepted as a problem shared by the members of that society. When the state recognizes its intervention, either to mitigate, reduce, or, at best, solve them, it transforms them into state problems. Examples of such problems may be air pollution, unemployment rates, or health problems, such as the COVID-19 pandemic. At this stage, the state identifies the root causes of the problem and finds alternatives to solve it. These are analytical actions aimed at understanding the problem and being able to propose actions to mitigate it, recording data, analyzing information, and considering the possible consequences of the intervention. AI through machine learning and expert systems could reveal patterns of needs and hidden problems in society. This could also anticipate decision making. Furthermore, AI could provide data to settle agenda decisions, and to justify actions and decision-making.

Policy formulation is part of the pre-decision phase of policymaking. It involves identifying and/or developing a set of policy alternatives to address a problem and narrowing that set of solutions in preparation for the final policy decision. According to Cochran and Malone, cited in Sidney (2017, p. 79), "policy formulation focuses on the what: What is the plan for addressing the problem? What are the objectives and priorities? What are the options for achieving those objectives? What are the costs and benefits of each of the options? What externalities, positive or negative, are associated with each alternative?". After the state considers the evidence and the possible consequences of the intervention, it chooses from a series of alternatives the best option for its objectives, capabilities, and resources. At this stage, an action plan is drawn up that includes a series of activities to be carried out to attack the public problem, and participants in the public policy process have moved the problem to the political agenda. AI could provide some simulations about policies to assess viability or rejection. It could also help to improve previous decisions with machine learning algorithms and to expand or interrelate government decisions in several governmental layers.

Public policy implementation lies at the intersection of public administration, organizational theory, public management research, and political science studies (Pülzl & Treib, 2017). Government policies are actions that a government establishes to generate new transaction patterns or institutions or to change established patterns within old institutions. Policy, formulated by a government, serves as a force for change in society. As policies are implemented, those who implement them and those affected by them may experience benefits, but also tensions, pressures, and conflicts (Smith, 1973). In this stage, the actions

established in the formulation phase are implemented to achieve the policy objectives. It is here where the coordination of the different actors involved stands out, both in the transmission of information and the use of available resources, as well as the emerging events that were not considered and that influence policy implementation. AI at this stage monitors the advances, problems, and challenges of implementation, and, if it is the case, finds solutions and alternatives for emergent or sudden problems in implementation.

The *policy evaluation* involves various social sectors, which are affected by the results of public policy, its justification, and effectiveness. At this stage of the public policy cycle, citizens may ask themselves, among other questions that governments must consider and analyze: Is the payment of taxes justified? Does the government meet the expectations it promoted? Does it offer public services efficiently and fairly? Does it use public resources honestly and transparently? Is state intervention necessary? Does it solve or create problems? (Guerrero, 1995). The decisions taken by the state to reduce or solve the public problem and the defined goals are compared with the results and performance of the actions carried out. It also seeks to understand whether the effects of such intervention have solved the problem set out in the agenda-setting. AI assessment could be faster and provide new insights and different approaches for the same problem as a result of machine learning (ML). Implementing the assessment process with instant data could be faster using different tools (e.g., data mining, expert systems), identity patterns, and points alerts. AI will change the assessment process making it instant and permanent, not at the end of the policy cycle.

Revisiting the Public Policy-Cycle in the Age of Artificial Intelligence

The purpose of this section is to describe the general policy-cycle framework in the age of AI. In the twenty-first century, some decisions on public policies that determine the daily lives of citizens are not in legal regulations but are instead in AI programs created by scientists and innovators in private (and monopolistic) environments. The impact of deep learning and intelligent predictive models on public policy design and outcomes needs to be understood by all countries to recognize the potential benefits, as well as to mitigate risks in areas of overlap. Good policy can usher in AI governance for the benefit of humanity.

AI has the potential to change organizational processes and how activities are commonly performed (Engin & Treleaven, 2019; Fernandez-Cortez et al., 2020). The disruption of this type of intelligent technology allows large-scale automation, including decision-making (Valle-Cruz et al., 2020). The AI techniques implementation in the public sector enhances dynamism and agility in the public policy cycle, accelerating processes and activities. In this regard the speed of decision-making in the policy cycle became a variable that substantially affects its process, as it can be slow and overwhelmed by the large number of problems faced by governments (Criado, 2021; Jimenez-Gomez et al., 2020; Wirtz et al., 2019). Moreover, the capabilities developed to evaluate results and release information slow down the process when generating strategies based on public policies. With the implementation of AI techniques in the public sector, there is the potential to accelerate public policy processes and governance (Valle-Cruz et al., 2020).

The dynamism and agility in the public policy cycle aims to fill the information processing and decision-making gap between the government, the different productive sectors, and social actors (Gritsenko & Wood, 2020; Katzenbach & Ulbricht, 2019). The disruptive nature of AI can transform organizations and the political process. In this regard, AI can be a double-edged sword: first, as a tool for improving organizational efficiency, and second, as a mechanism to reinforce bad bureaucratic practices, such as artificial discretion (Young, Bullock, et al., 2019). Regarding the former, AI allows analyzing big data, finding patterns and information that humans cannot detect (Pencheva et al., 2020). Also, AI can perform routine tasks with greater accuracy and without interruption and can perform high-risk or heavy-lifting activities without exposing human life. AI implementation in the public sector provides the freedom to do creative or innovative practices strategical to decision making. Also, AI can enable humans to take advantage of the benefits it provides to enhance their capabilities, as well as to generate a framework for social inclusion in the policy cycle through methodologies that simplify complexity and democratize access to co-production mechanisms (Sun & Medaglia, 2019; Wirtz et al., 2020).

The policy cycle in the age of AI follows a dynamic and flexible flow, rather than the traditional conception of a single cycle, which requires several iterations to produce a public policy. AI can generate an evolution of the traditional cycle into an incremental and spiral approach, allowing feedback at each stage of the cycle by analyzing data with AI techniques and simulations (Valle-Cruz et al., 2020). Consequently, it will not be necessary to wait for the implementation phase to have results for evaluation. AI can produce scenarios, simulations, and different solutions for each stage of the public policy cycle in near real-time. In the dynamic policy cycle, agenda-setting can happen at the same time as policy evaluation; policy formulation and decision making could occur almost at the same time as policy implementation (e.g., it can be caused by emerging or unknown situations such as the COVID-19 pandemic). AI applications valid for the public policy cycle can monitor data to discover patterns helpful in predicting events. However, some risks, such as algorithmic discrimination, opacity, social control, human replacement, data privacy, and the increasing digital divide, should be ruled by AI governance. AI has had a forced debut with the pandemic. The efforts to produce a vaccine, understand the virus, manage the dissemination, and population control will create a synergy for all AI developments.

A contemporary problem is the COVID-19 pandemic, in which governments around the world had to improvise policies and strategies to combat the potential harms of the virus (Yeh & Cheng, 2020; Yousefpour et al., 2020). The stakeholders during the outbreak consisted of politicians, private companies, and members of civil society who had as nodal points to fight the pandemic, detect the virus, and avoid deaths and the spread of the virus, as well as avoid economic collapse (Baxter & Casady, 2020; Storr et al., 2021). Some established and emerging norms had to do with the measures of healthy distance, the closing of some businesses, and the ethical and equitable use of the technologies and innovations put in place during the pandemic (Vaishya et al., 2020).

As an emerging global problem, some processes had not been fully established, but digitization, emergent decision-making, risk avoidance, and response mechanisms were recurring processes during the pandemic. In this regard, Barbieri et al. (2021) analyzed relevant papers that directly address the AI adoption and new technologies in the management of pandemics and communicable diseases such as SARS-CoV-2. Three of the examples identified by Barbieri and colleagues are: (1) the Center for Systems Science and Engineering

at Johns Hopkins University created a digital platform for sharing data related to the worldwide spread of the COVID-19 pandemic; (2) several mobile apps were developed in an attempt to limit the growing number of COVID-19 cases and deaths, decreasing misinformation and confusion, tracking COVID-19 symptoms and the mental health of citizens, decision-making, contact tracing, home monitoring, and isolation; and (3) in addition to using AI algorithms to the personalization of medical practice, assistance in early detection of patient's shock and septic conditions of COVID-19 infections, and forecast capabilities in order to assess the spread of the pandemic.

The Artificial Intelligence in the Public Policy Cycle during the Pandemic

This section aims to analyze the pandemic through the lens of AI in the public policy cycle. The COVID-19 emergency is an exceptional situation that made it possible to understand how public policies would be empowered by AI, because, among other issues, technology was a real option to appease the pernicious effects of the pandemic. In this regard, the COVID-19 pandemic was a trigger for the implementation of AI-supported public policies to diminish the contagion (Chamola et al., 2020), assist physicians in diagnosis (Barbieri et al., 2021), construct ML-based models to forecast virus-spreading (Bastos & Cajueiro, 2020), use robots and drones to reinforce the healthy distance measures (Bogue, 2020), as well as pattern recognition to detect SARS-CoV-2 in chest X-rays (Brunese et al., 2020). To analyze this case study, the authors considered the four phases of the AI-powered public policy cycle to combat the COVID-19 pandemic: agenda setting, policy formulation and decision making, policy implementation, and policy evaluation.

Agenda setting

Public policy agenda items are imposed by the ideology of the moment or by reality. In this case, the health emergency had an essence of improvisation and innovative proposals to combat the pandemic supported by the collaboration of private sector and civil society. In the face of an unprecedented event, no known path would provide a solution to the problem that arose—it was necessary to learn and develop human resources and health specialists on the progress of the pandemic. However, the health emergency establishes five agenda nodal points that are described in the following sections: (1) detection and monitoring of cases of COVID-19; (2) isolation of cases; (3) hospitalization, recovery, or deaths; (4) dissemination of hygiene and social distancing measures; and (5) vaccination processes.

Policy formulation and decision making

Policy formulation and decision-making had to be agile as the virus spread around the world. The private sector played a pivotal role in providing technologies to combat the

pandemic and to boost the collaboration with governments immersed in and emerging and improvising decision-making. The following discussion is a description of the established policies' parameters during the pandemic.

Detection and monitoring of COVID-19 cases

In the detection and monitoring of COVID-19 cases, the decision parameters of this public policy were fourfold: (1) detection of the COVID-19 virus through rapid tests or PCR tests; (2) detection of the virus directly in hospitals or health centers; (3) symptom detection through telemedicine, drones, or smart glasses; and (4) no virus detection. As for the formulation of case monitoring, the decision parameters evolved: (1) monitoring from AI tools, such as Blue Dot technology (Allam et al., 2020; Tuite et al., 2020); (2) monitoring people's mobility through Big Data and ICTs; and (3) monitoring trends through the measurement of immunization data, deaths, and cases by zones. The case of the Netherlands is an example of this stage, referring to the formulation of public policies immersed in an environment with limited information and facing many uncertainties towards adaptive governance (Janssen & Van Der Voort, 2020). The use of AI tools, such as predictive analytics, forecasted how many confirmed COVID-19 cases and deaths can be expected in the near future. Alsinglawi et al. (2021) used this technique to create simulations from data extracted from the Johns Hopkins University Centre for Systems Science and Engineering, and they produced some algorithms to forecast and monitor the pandemic.

Isolation of cases

The formulation of a case isolation policy has been very controversial. For instance, from the lockdown and cessation of economic activities in China, to partial closures or specific measures, to the spread of the virus to other countries, such as Italy, Spain, the US, and Mexico, among others. The formulation of the public policy can be recognized in five phases: (1) closure of productive activities, schools, and offices; (2) partial closure of productive activities for hours or with a reduction in the capacity of people; (3) partial closure of productive activities and offices but the opening of schools; (4) gradual opening of activities with sanitary measures; and (5) letting economic activity flow regardless of the increase in contagions.

Hospitalization, recovery, or deaths

With the advance of the pandemic and the collapse of the health care system in several countries, in addition to the global economic downturn, measures were taken for hospitalization, recovery, and death. The following measures were taken for the formulation of this policy: (1) isolation of specialized COVID-19 hospitals; (2) isolation of areas, wards, or zones of specialized hospitals; (3) case reception areas, isolated from the rest; (4) vaccination of hospital medical and administrative personnel; and (5) monitoring of hospital data. Bari and Coffee (2020) designed an experimental AI tool to find patterns to "determine which mildly ill patients were likely become severely ill"; however, this AI tool was trained with a small number of data and required more data to increase its accuracy.

Dissemination of hygiene and social distancing measures

For the dissemination of social distancing and hygiene measures, social networks and mass media were mostly used. There are cases in which ML and Big Data techniques were used to enhance telemedicine (PAHO, 2020), in addition to the use of robots and drones to keep a healthy distance and prevent medical personnel from being in direct contact with the infected (Laguipo, 2020). However, public policies were established as follows: (1) prioritize official pandemic information channels using technology; (2) discard false or erroneous information ("fake news", conspiracy theories) about the pandemic; (3) daily or hourly publication of information relevant to pandemic control; and (4) identify the disseminators or propagators of false information to neutralize them. As a complement, Janeja (2020) proposes the use of deep learning algorithms to run artificial neural networks and find patterns on unstructured data, that comes from several sources, such fake news, and provides some help to prioritize information.

Vaccination process

This public policy was decided according to each country. Therefore, the decision parameters for promoting citizens' vaccination, in addition to the type or brand of vaccine, can be very different, such as the following: (1) vaccination by age groups and by regions of the country; (2) vaccination is open to any person and they can acquire the brand of their preference; (3) vaccination is regulated and administered by the government in public institutions open to the public; and (4) the type or brand of vaccine being administered. An example of AI is the project of Keriwal (2021), in which they found that processing geo-located Twitter data using readily available ML techniques could more accurately predict vaccine hesitancy by ZIP code.

Policy implementation

The implementation of public policies was an urgent issue in the nations. Although there are different areas of health public policy, this discussion presents those that were supported by AI techniques and data exploitation. The implementation of public policies is described using the available technological tools.

Detection and monitoring of cases of COVID-19

This public policy was established in a differentiated manner. Some countries did mass testing of their citizens to determine the extent of the virus and create models. Countries such as Singapore, Malaysia, and Germany (Tay, 2021) carried out these actions and relied on mapping technologies, such as Blue Dot technology. The monitoring of cases was done in the first instance by the mobility of cellular telephony, through companies such as Google and Apple (Apple, 2021; Google, 2021). Later, detection and monitoring migrated to registration platforms with QR codes in shopping malls and accesses (Rahman et al., 2021).

Isolation of cases

Forced closure, lockdown, and the reduction of daily activities were some of the other public policies implemented. To enforce this, countries such as Spain promoted the use of drones to monitor their cities (Chamola et al., 2020). Singapore promoted the use of robots to monitor parks (Statt, 2020). China developed mass surveillance algorithms for its video cameras to measure the distance and facial recognition of its citizens who were outside and would have been forced to lock down to prevent the spread of the virus. Much of the implementation of this type of technology was driven by the private sector for use in government. Other countries went further and monitored the progress of citizens carrying the virus using combined social networking and AI technologies to drive mass surveillance, as in the case of China and Poland (Malgorzata, 2020).

Hospitalization, recovery, or deaths

For hospitalization processes, several actions were taken according to the possibilities of each country and technologies provided by the private sector. Telemedicine was used to relieve the burden on hospitals. Robots at the entrance of hospitals took vitals and temperatures. Robots carried laboratory samples. Smart Glasses were used to identify symptomatic people. AI-based evidence was used to identify people infected with the SARS-CoV-2 virus.

Dissemination of hygiene and social distancing measures

In the case of the dissemination of hygiene and social distancing measures, they were carried out in various ways using technology. Examples include COVID-19 passports at airports; dissemination of information in citizens' social networks; social distancing control monitored by CCTV cameras in cities; social distancing implemented by robots; the identification of fake news and the creation of bots to neutralize them; dashboards of information on the progress of the virus, such as the one used by Johns Hopkins to measure the progress of the pandemic; and the use of thousands of predictive algorithms to determine the spread of the pandemic.

Vaccination process

The vaccination process according to the public policy decision had several types of implementations, including the following: website registration of citizens to receive their vaccination, in-person registration at the vaccination centers, and registration using QR code. Here, it was crucial to identify the regions with the greatest economic activities and specific needs to select the most vulnerable and important sectors to be vaccinated. Some countries, especially developing countries, had to distribute a minimum number of vaccines for millions of inhabitants.

Policy evaluation

Although the COVID-19 pandemic is still ongoing at the time of writing (the 4th wave), some public policies have been evaluated; however, most technological implementations

have not because they are in the testing phase. Heaven (2021), quoting the Turning Institute report, states that many hundreds of predictive tools using AI were developed, but none of them made a real difference and some are potential harmful. Also Wynants et al. (2020) reviewed 232 algorithms for diagnosing patients and they found that none of them were fit for clinical use. Heaven (2021) points out, "[t]he pandemic was a big test for the AI . . . it creates an unrealistic expectation to use tools that help overcome the pandemic."

In the case of policy evaluation, some measures of implemented policies are related to the use of epidemiological traffic lights through data concentration; this is one way to measure the effectiveness of health policies, despite the use of technology. On the other hand, the economic growth allows us to know how much the massive closures of businesses and the measures of healthy distance have affected them. Another type of indicator that helps to understand the impact of policies is mortality and vaccination rates. However, there is a gap of tools—using AI—to effectively measure the impact of public policies in real time and without mistakes.

Final Comments and Conclusions

This chapter aimed to present a review of artificial intelligence on the public policy cycle in the times of COVID-19. In the case of COVID-19, the AI policies were related, predominantly, to innovation boosted by the private sector. The SARS-CoV-2 virus generated disruptions in most aspects of society worldwide. Therefore, policy and decision-making, within governments, was improvised, with emergent solutions to solve the problems that had arisen. Throughout the public policy cycle, the emerging problem was to combat COVID-19 and all its health and economic consequences as nodal points. Some rules identified are related to access to data, the privacy of information, as well as restriction measures and healthy distance supported by smartphones, robots, and drones.

Since the coronavirus infection cases skyrocketed, there has been a growing number of initiatives to help in the fight against the pandemic that threatened a deep economic recession. The outbreak triggered massive demand for digital health solutions. However, when fighting a pandemic almost nothing matters more than speed, which is why public healthcare policies had to be implemented in an emergent manner. In this regard, the authors also found aspects related to the interaction between the citizen and the government, as well as the human resources development, with emergent measures against the pandemic. Some projects show a strong public–private partnership, as well as some aspects related to information privacy. The following paragraphs explain each stage of the public policy cycle according to COVID-19 case analysis.

Agenda setting began with the participation of actors from the private sector, members of civil society, and the government. Because this was a little-known problem, the essence of the agenda was innovation fostered by the private sector for the norms and processes design. Much of the AI technology was provided by the private sector as governments were overwhelmed by the needs generated by the pandemic. Health personnel had to be prepared to face the virus, so there was a great development of human resources in the face of the emergency. At this stage, the planning of the agenda began with equity, accountability, data privacy, and information efficiency; government-to-citizen interaction, decentralized

administration, and networking were not identified. At this stage, there were no established standards due to the emerging and unknown situation.

In the COVID-19 era, policy formulation and decision making is a stage impregnated with innovation *and* public–private partnership. Participation and equity are essential to norms and processes providing solutions to the pandemic problem. In addition to the importance of policy formulation, it was essential to consider vaccination campaigns and measures of protection and healthy distance with a sense of equity. Policies have been formulated to monitor and track infected persons. For this reason, it is necessary to preserve data privacy and information efficiency. Government-to-citizen interaction is necessary for the efficient formulation of policies. Human resource development is still necessary at this stage of the public policy cycle. Decentralized administration and networking were not identified.

The problem and nodal points are no longer analyzed in the policy implementation because they were established and analyzed in the previous stages. At this stage, is important to encourage the interaction between actors, the respect of norms, and the implementation of processes to combat the pandemic. The private and public sectors must be held accountable for AI-based information and processes. The public–private partnership, government-to-citizen interaction, and equity are fundamental to policy implementation. Participation, decentralized administration, networking, and human resource development were not identified at this stage.

In the final stage of the public policy cycle, accountability for the norms and processes carried out during the pandemic is critical. As the spread of the SARS-CoV-2 virus has not ended, innovation governs the processes of the policy cycle. Public policy evaluation is based on epidemiological traffic lights, vaccination, and death rates, as well as GDP growth.

In brief, on the agenda setting public policy, we learned that more development on AI algorithms could help understand the COVID-19 spread and prevent more deaths, and could generate alternatives for decision-making that would help global leaders attend to this problem. The second stage of public policy—policy formulation—shows that the lack of ethical criteria, the absence of regulations for AI development, and the lack of control of private companies on the development of solutions for government cause risks and challenges for the proposed solutions to face the pandemic's evolution. In this race to solve problems and difficulties, the improvised solutions create indirect consequences. This is the reason to promote a global governance framework on AI and the public sector to overcome this problem soon. The third stage of public policy—implementation—includes shadows and lights. On the lights, the fast development and performance of AI solutions such as apps, software, programs, vaccines, and tests face the immediate problems caused by COVID-19. Implementation also creates the opportunity to foster government and industry relations and promote many AI developments, some of which work while some others do not. The implementation fosters public innovation, human resources development in emergencies, and helps to manage crises. In the shadows, many of these developments could have more effect or trials if more government money is invested in AI projects.

The fourth stage is policy evaluation. This stage is an ongoing process. The information to assess AI policies in the COVID-19 emergency is scarce and ambiguous. Some reports show important advances, and some others offer very few advances.

The pandemic has served as an experience for governments to consider AI as a mechanism that can provide potential benefits. Firstly, in terms of innovation and efficiency, and secondly, for the improvement of interaction between governments and citizens and the

development of human resources empowered by AI and towards AI. Thus, the great potential of AI has arisen from the collaboration between different actors: citizens, public and private sector, NGOs, and the exploitation of data, for the solution of various problems that afflict society.

We do not know the future and scope of AI, and at this point we only have pure speculation. However, the future of the public policy cycle synergized by AI has a tremendous impact we expect to see, for example:

An improved agenda setting revolutionized to be faster and more efficient, with fewer data errors and including more variables, such as gender, age, health, literacy, and weather changes.

A policy formulation that contains lessons learned from previous experiences (deep learning), resulting from a rigorous analysis of alternatives and not only from political and ideological experience, relying more on data and simulations and less on political ideologies.

Faster policy implementation and closer coordination and collaboration among government agencies resulting from immediate, high-quality, and simultaneous data sharing. Accurate monitoring of progress in implementing decisions and possible alerts of risks or threats to risks through machine learning, pattern identification, and simultaneous surveillance of other variables and processes.

In the policy evaluation, we expect to interrelate automated learning, knowledge of computational errors, and human experience, including immediate feedback that is linked to very clear and concrete projects and processes.

Thus, the ideas that have emerged on issues of governance and public policy, continuous improvement, accountability, open data, efficiency, government–citizen interaction, and public value can be harnessed with the use of AI. The COVID-19 pandemic highlighted the importance of AI governance in the public policy cycle. AI has the potential to streamline decision-making in the public policy process, fostering participation, equity, accountability, privacy, and innovation.

References

Allam, Z., Dey, G., & Jones, D. S. (2020). Artificial intelligence (AI) provided early detection of the coronavirus (COVID-19) in China and will influence future Urban health policy internationally. *AI 1*(2), 156–165.

Alsinglawi, B., Mubin, O., Alnajjar, F., Kheirallah, K., Elkhodr, M., Al Zobbi, M., . . . & Dev, K. (2021). A simulated measurement for COVID-19 pandemic using the effective reproductive number on an empirical portion of population: epidemiological models. *Neural Computing and Applications*, 1–9. https://doi.org/10.1007/s00521-021-06579-2

Apple. (2021). *COVID-19—Mobility Trends Reports Online*. https://covid19.apple.com/mobility.

Bari, A., & Coffee, M. (2020). We designed an experimental AI tool to predict which COVID-19 patients are going to get the sickest. *The Conversation*. https://theconversation.com/we-designed-an-experimental-ai-tool-to-predict-which-covid-19-patients-are-going-to-get-the-sickest-136125

Barbieri, D., Giuliani, E., Del Prete, A., Losi, A., Villani, M., & Barbieri, A. (2021). How artificial intelligence and new technologies can help the management of the COVID-19 pandemic. *International Journal of Environmental Research and Public Health 18*(14), 7648.

Bastos, S. B., & Cajueiro, D. O. (2020). Modeling and forecasting the early evolution of the COVID-19 pandemic in Brazil. *Scientific Reports 10*(1), 1–10.

Baxter, D., & Casady, C. B. (2020). Proactive and strategic healthcare public–private partnerships (PPPs) in the coronavirus (COVID-19) epoch. *Sustainability 12*(12), 5097.

Birkland, T. A. (2017). Agenda setting in public policy. In F. Fischer and G. Miller (Eds.), *Handbook of public policy analysis* (pp. 89–104). Routledge.

Bogue, R. (2020). Robots in a contagious world. *Industrial Robot: The International Journal of Robotics Research and Application*, *47*(5), 673–642. https://doi.org/10.1108/IR-05-2020-0101

Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Machine learning for coronavirus COVID-19 detection from chest x-rays. *Procedia Computer Science 176*, 2212–2221.

Campion, A., Gasco-Hernandez, M., Jankin Mikhaylov, S., & Esteve, M. (2020). Overcoming the challenges of collaboratively adopting artificial intelligence in the public sector. *Social Science Computer Review*, *40*(2), 462–477. https://doi.org/10.1177/0894439320979953

Cantú, C., Cheng, G., Doerr, S., Frost, J., & Gambacorta, L. (2020). On health and privacy: technology to combat the pandemic. *BIS Bulletin*,. *17*, 1–2. https://www.bis.org/publ/bisbull17.pdf

Capucha, L., Nunes, N., & Calado, A. (2020). Artificial intelligence as a resilient tool for fighting inequalities in the COVID-19 crisis. *Artificial Intelligence as a Resilient Tool for Fighting Inequalities in the COVID-19 Crisis 2*, 10–19.

Cath, C. (2018). *Governing artificial intelligence: ethical, legal and technical opportunities and challenges*. The Royal Society Publishing.

Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. *IEEE Access 8*, 90225–90265.

Churchman, C. West (1967). Free for All. *Management Science*, *14*(4): B-141–B-146. https://doi.org/10.1287/mnsc.14.4.B141

Coveri, A., Guarascio, D., & Landesmann, M. (2020). International production, structural change and public policies in times of pandemics. *Journal of Industrial and Business Economics 47*(3), 363–369.

Criado, J. I. (2021). Inteligencia artificial (y administración pública). *EUNOMÍA. Revista En Cultura de La Legalidad 20*, 348–372.

Criado, J. I., Valero, J., & Villodre, J. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity 25*(4), 449–470.

Dafoe, A. (2018). *AI governance: A research agenda.* Governance of AI Program, Future of Humanity Institute, University of Oxford. Oxford, UK.

Engin, Z., & Treleaven, P. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *Computer Journal 62*(3), 448–460. https://doi.org/10.1093/comjnl/bxy082.

Fernandez-Cortez, V., Valle-Cruz, D., & Gil-Garcia, J. R. (2020). Can artificial intelligence help optimize the public budgeting process? Lessons about smartness and public value from the Mexican federal government. *2020 Seventh International Conference on EDemocracy & EGovernment (ICEDEG)*, 312–315. https://doi.org/10.1109/ICEDEG48599.2020.9096745.

Malgorzata, G. (2020). *COVID-19: Poland launches an official tracking app*. On *Squire Patton Boggs*. https://www.ceelegalblog.com/2020/04/covid-19-poland-launches-an-official-tracking-app/

Google. (2021). *Community Mobility Report*. https://www.google.com/covid19/mobility/

Gritsenko, D., & Wood, M. (2020). Algorithmic governance: A modes of governance approach. *Regulation & Governance*, *6*(1), 45–62.

Guerrero Amparán, J. P. (1995). La evaluación de políticas públicas: enfoques teóricos y realidades en nueve países desarrollados. Gestión y Política Pública IV(1), 47–115.

Heaven (2021)Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. Online: https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

Janeja (2020) Predicting the coronavirus outbreak: How AI connects the dots to warn about disease threats. *The Conversation*. Online: https://theconversation.com/predicting-the-coronavirus-outbreak-how-ai-connects-the-dots-to-warn-about-disease-threats-130772

Jann, W., & Wegrich, K. (2017). Theories of the policy cycle. In F. Fischer and G. Miller (Eds.), *Handbook of public policy analysis* (pp. 69–88). Routledge.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly 37*(3), 101493.

Janssen, M., & Van Der Voort, H. (2020). Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic. *International Journal of Information Management 55*, 102180.

Jimenez-Gomez, C. E., Cano-Carrillo, J., & Lanas, F. F. (2020). Artificial intelligence in government. *IEEE Annals of the History of Computing 53*(10), 23–27.

Jr, E. A., Fornaciali, M., Batista, A., Gazzola, M., da Silva, L. P., Patrão, D. F. C., & Jr, M. F. (2020). Use of artificial intelligence in health. *Internet Sectoral Overview*, *12*(2), 1–11

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review 8*(4), 1–18.

Keriwal, M. (2021). Matching tweets to ZIP codes can spotlight hot spots of COVID-19 vaccine hesitancy. In *The Conversation*. https://theconversation.com/matching-tweets-to-zip-codes-can-spotlight-hot-spots-of-covid-19-vaccine-hesitancy-169596

Knoepfel, P., Larrue, C., & Varone, F. (2003). Las políticas públicas. In P. Knoepfel (Ed.). *Análisis y conducción de las políticas públicas* (pp. 25–37). U. de Monterrey.

Laguipo, A. B. B. (2020). Robots and drones can help curb the coronavirus pandemic. *News Medical Life Sciences Online*. https://www.news-medical.net/news/20200326/Robots-and-drones-can-help-curb-the-coronavirus-pandemic.aspx

Lahera, E. (2004). *Capital institucional y desarrollo productivo: un enfoque de pol{\'\i}ticas públicas*. CEPAL.

Li, S., Li, H., Song, H., & Chen, M. (2021). Mitigating tourism social costs during a pandemic: evaluating residents' perceptions and behavioral intentions. *Journal of Travel Research*, *61*(3), 493–510. https://doi.org/10.1177/00472875211000447.

PAHO. (2020). *Rehabilitation considerations during the COVID-19 outbreak*. PAHO Washington, DC.

Paquet, M., & Schertzer, R. (2020). COVID-19 as a complex intergovernmental problem. *Canadian Journal of Political Science/Revue canadienne de science politique*, *53*(2), 343–347. https://doi.org/10.1017/S0008423920000281.

Park, J., & Chung, E. (2021). Learning from past pandemic governance: Early response and public–private partnerships in testing of COVID-19 in South Korea. *World Development 137*, 105198.

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big data and AI—A transformational shift for government: So, what next for research? *Public Policy and Administration 35*(1), 24–44.

Pülzl, H., & Treib, O. (2017). Implementing public policy. In F. Fischer and G. Miller (Eds.), *Handbook of public policy analysis* (pp. 115–134). Routledge.

Rahman, M. T., Shuva, T. F., Khan, R. T., & Nasir, M. K. (2021). A review of contact tracing approaches for controlling COVID-19 pandemic. *Global Journal of Computer Science and Technology*, *21*(1), 31–38.

Rodríguez-Rodríguez, I., Rodríguez, J.-V., Shirvanizadeh, N., Ortiz, A., & Pardo-Quiles, D.-J. (2021). Applications of artificial intelligence, machine learning, big data and the internet of things to the COVID-19 pandemic: A scientometric review using text mining. *International Journal of Environmental Research and Public Health 18*(16), 8578.

Sarkar, S. (2021). Breaking the chain: Governmental frugal innovation in Kerala to combat the COVID-19 pandemic. *Government Information Quarterly 38*(1), 101549.

Sidney, M. S. (2017). Policy formulation: design and tools. In F. Fischer and G. Miller (Eds.), *Handbook of public policy analysis* (pp. 105–114). Routledge.

Smith, T. B. (1973). The policy implementation process. *Policy Sciences 4*(2), 197–209.

Statt, N. (2020). Boston Dynamics' Spot robot is helping hospitals remotely treat coronavirus patients. *The Verge*. https://www.theverge.com/2020/4/23/21231855/boston-dynamics-spot-robot-covid-19-coronavirus-telemedicine.

Storr, V. H., Haeffele, S., Lofthouse, J. K., & Hobson, A. (2021). Entrepreneurship during a pandemic. *European Journal of Law and Economics*, *1*(1) 1–23. https://doi.org/10.1007/s10657-021-09712-7.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly 36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008.

Tay, S. (2021). *Could apps help India manage its second Covid surge?* GovInsider. https://govinsider.asia/digital-gov/could-apps-help-india-manage-its-second-covid-surge-neeta-verma-nic/.

Thierer, A. D., Castillo, A., & Russell, R. (2017). *Artificial intelligence and public policy*. Mercatus Research Paper, Available at SSRN: https://ssrn.com/abstract=3021135 or http://dx.doi.org/10.2139/ssrn.3021135

Tuite, A. R., Bogoch, I. I., Sherbo, R., Watts, A., Fisman, D., & Khan, K. (2020). Estimation of coronavirus disease 2019 (COVID-19) burden and potential for international dissemination of infection from Iran. *Annals of Internal Medicine 172*(10), 699–701.

Vaishya, R., Haleem, A., Vaish, A., & Javaid, M. (2020). Emerging technologies to combat the COVID-19 pandemic. *Journal of Clinical and Experimental Hepatology 10*(4), 409–411.

Valle-Cruz, D., Criado, J. I., Sandoval-Almazán, R., & Ruvalcaba-Gomez, E. A. (2020). Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Government Information Quarterly 37*(4), 101509.

Valle-Cruz, D., Fernandez-Cortez, V., & Gil-García, J. R. (2021). From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. *Government Information Quarterly*, *39*(2). https://doi.org/https://doi.org/10.1016/j.giq.2021.101644

Villanueva, L. F. A., Luis, F., & Porrua, M. A. (2007). *Problemas públicos y agenda de gobierno*. Porrúa.

Warren, A., & Hillas, A. (2020). Friend or frenemy? The role of trust in human-machine teaming and lethal autonomous weapons systems. *Small Wars & Insurgencies 31*(4), 822–850.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration 42*(7), 596–615.

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration 43*(9), 818–829.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E. et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal *BMJ*, *369*: m1328. https://doi.org/10.1136/bmj.m1328.

Yeh, M.-J., & Cheng, Y. (2020). Policies tackling the COVID-19 pandemic: A sociopolitical perspective from Taiwan. *Health Security 18*(6), 427–434.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance 2*(4), 301–313.

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2019). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4*(3), 244–258.

Yousefpour, A., Jahanshahi, H., & Bekiros, S. (2020). Optimal policies for control of the novel coronavirus disease (COVID-19) outbreak. *Chaos, Solitons & Fractals 136*, 109883.

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3). https://doi.org/10.1016/j.giq.2021.101577.

SECTION VI

AI GOVERNANCE FROM THE GROUND UP (VIEWS FROM THE PUBLIC, IMPACTED COMMUNITIES, AND ACTIVISTS WITHIN THE TECH COMMUNITY)

BAOBAO ZHANG

CHAPTER 28

PUBLIC OPINION TOWARD ARTIFICIAL INTELLIGENCE

BAOBAO ZHANG

Introduction

Public opinion toward artificial intelligence (AI) has become an emerging area of study within AI policy. Much of the existing survey research has been conducted by companies, think tanks, and governments rather than academics. Nevertheless, the growth of AI ethics as a field of study has increased the number of academic publications on the topic. Furthermore, survey work has expanded beyond high-income countries to include respondents in the Global South. This chapter synthesizes and discusses research on public opinion toward AI, while at the same time proposing new directions for research.

There are two chief reasons why understanding public opinion toward AI matters for AI governance. First, from a normative perspective, the public is a major stakeholder in shaping the future of AI and, therefore, should be included in discussions around AI governance. Secondly, in cases of other technologies (e.g., nuclear energy and genetically modified foods), the public could shape the development and deployment of AI. Understanding what they think about AI will help us anticipate future political contestation and consumer behavior.

Even as AI systems become more widely deployed in public, most of the existing works in AI ethics focus on ethics principles developed by tech companies, governments, think tanks, or academic institutions (Fjeld et al., 2020). Much of the work is prescriptive: they describe what ideal "ethical" AI systems should look like. These principles, such as respecting human rights or preserving privacy, aim to protect the public's welfare. Nevertheless, public input to shape these ethical principles has been limited or costly. Participatory design has been suggested as one way to ensure that AI systems work in the public's interest (Kulynych et al., 2020). While public opinion research may not directly impact computer scientists' design choices, it can help inform policymakers and tech companies of the public's concerns about AI in general or specific applications of AI. For instance, the research could illuminate the concerns that are not salient in elite discourses about AI governance or reveal consensus around an ethics principle.

Furthermore, survey research could illuminate how the public has conflicting views about what they consider ethical uses of AI. Indeed, some of the published AI ethics principles conflict with each other (Whittlestone et al., 2019). For instance, the lack of non-white people in image databases used to train computer vision systems produces less accurate predictions for darker-skinned individuals. Nevertheless, gathering additional training data could mean greater surveillance of populations that are overly policed (Cooper & Abrams, 2021). These theoretical tensions also play out in disagreements between survey respondents. In the US, support for facial recognition technology varies significantly by race, party identification, and age (Smith, 2019). The Moral Machine project, which collected 40 million decisions in 233 countries and territories, reveals that respondents in different regions and cultures have divergent preferences regarding autonomous vehicles' behavior in moral dilemmas (Awad et al., 2018). This type of research necessarily complicates AI ethics by revealing that consumers and voters disagree about how AI systems should be developed and deployed.

As seen throughout history and in contemporary events, the public has shaped technology policies, including genetically modified (GM) foods, nuclear power, and vaccines. Understanding public sentiments could help policymakers, activists, and tech companies anticipate mass mobilization around AI-related issues, particularly around calls to ban specific AI applications. First, they act as direct consumers and boycott products or services produced by the technology. For instance, European consumers perceive GM foods as risky and have been reluctant to consume them (Frewer et al., 2004). In the past two decades, vaccine confidence has declined in many parts of the world (de Figueiredo et al., 2020), leading to stagnating vaccination rates (Requejo et al., 2020). Second, the public can demand change in regulation through mass mobilization, activism, and voting. For instance, anti-vaccine groups in the US have pushed states to adopt non-medical exemptions for vaccines (Olive et al., 2018). Voters opposing nuclear power have voted in referendums to ban (Austria in 1978; Italy in 2011) or phase out (Switzerland in 2017) nuclear power (Pelinka, 1983; Moody, 2011; BBC, 2017).

Those opposed to what they perceive to be unethical AI-adjacent technologies have adopted similar tactics to limit or ban the use of these technologies. For instance, organizations like the Electronic Frontier Foundation have discouraged consumers from purchasing Amazon's Ring home security system to protect user privacy and prevent excessive police surveillance (Guariglia, 2020). Furthermore, members of the public have increasingly protested against AI-adjacent applications deployed by governments. In 2020, hundreds of students in the UK protested outside the Department for Education, decrying an algorithm that predicted their exam grades and was unfair to students from lower socioeconomic backgrounds (Kolkman, 2020). As a result, officials reversed course and threw out the grades predicted by the algorithm. As AI and AI-adjacent applications become more widely deployed, consumer and citizen mobilization may become more widespread.

Here, I lay out the structure of this chapter. First, I summarize survey findings regarding the public's knowledge of AI and general trust in AI. I break down these results by country and by demographic subgroups, including gender and socioeconomic status. Secondly, I consider public attitudes toward four specific applications of AI: facial recognition technology, personalization algorithms, lethal autonomous weapons, and workplace automation. These four applications currently have high political salience around the globe; as a result, there exists a large trove of survey data on these topics. Finally, I conclude this

chapter by discussing four topics for further research: institutional trust in actors behind AI systems, the impact of experience and knowledge on attitudes toward AI, heterogeneity in attitudes toward AI, and how beliefs about AI impact consumer and civic behavior.

Fundamental Public Opinion Research: Knowledge and Trust

The first section of this chapter focuses on the public's knowledge about AI and their trust in AI systems in general. Existing research suggests that the public around the world is aware of AI, but their knowledge does not necessarily reflect current scientific understanding. The public's trust in AI differs greatly by geography and demographic characteristics.

Knowledge about AI

Two of the most fundamental questions in studying public attitudes toward AI include how much the public knows about the technology and how they define it. While the public may not have complete technical knowledge of AI, a 2018 survey conducted in eight low-, middle-, and high-income countries revealed that the vast majority of respondents have heard of AI (Kelley et al., 2019). Furthermore, this and other studies have demonstrated that the public has at least partial knowledge of what AI is by describing machines making decisions typically made by humans or mentioning related technologies like robots (Cave et al., 2019). While computer scientists tend to define AI by its technical functionality, the public's definition of AI tends to compare it with human behavior or intelligence. Furthermore, the public perceives AI as futuristic rather than something that they already interact with. These trends may be driven by how the news media and popular media depict AI systems. Recent criticism of existent algorithms and AI applications may change the public's understanding of the technology.

The canonical textbook *Artificial Intelligence: A Modern Approach* defines AI as "the study of agents that receive precepts from the environment and perform actions," where "each such agent implements a function that maps precept sequences to actions" (Russell & Norvig, 2020). A 2019 survey of AI and machine learning (ML) researchers found that 72 percent of respondents preferred definitions of AI that emphasized mathematical problem solving and technical functionality over definitions that compared machines with humans (Krafft et al., 2020). Although computer scientists prefer to define AI without emphasizing comparisons with humans, popular understanding of the technology tends to anthropomorphize AI (Salles et al., 2020). For instance, in a 2018 nationally representative survey in the UK, 42 percent of respondents "referred to computers performing tasks that replicated aspects of human cognition" and 25 percent referred to robots when describing AI (Cave et al., 2019). In a 2018 nationally representative survey in the US, respondents were more likely to label tech applications that can socially interact with humans (e.g., virtual assistants, social robots) as AI than applications that cannot (e.g., Google Translate, Google Search; Zhang & Dafoe, 2019).

Why the public anthropomorphizes AI could be explained by the media they consume. Across eight countries, the three top ways that the public reported learning about AI are through social media, TV reports and commentaries, and movies or TV shows (Kelley et al., 2019). These sources are likely to compare AI with human intelligence. A majority of English-language policy documents published by governments, tech companies, and civil society groups that defined AI did so in comparison to human cognition or behavior (Krafft et al., 2020). A review of AI in fictional narratives finds that AI is frequently depicted as intelligent machines embodied in humanoid forms (Cave et al., 2018). AI ethicists worry that representing AI as human-like could lead the public to be misinformed about the risks and benefits of AI (Cave et al., 2018). One particular concern is that the public fails to recognize how AI systems are deployed today, most of which are not humanoid robots but software embedded in commonplace sociotechnical systems (Krafft et al., 2020).

Related to the issue of anthropomorphizing AI, the public appears to be subject to the "AI effect." The AI effect is the phenomenon in which people perceive an AI system not to be "truly intelligent" once it solves a problem; as a result, AI is viewed as a futuristic technology rather than an existent one (McCorduck, 2004). Content analysis of open-ended responses from the public in eight countries revealed that 24 percent described AI as "futuristic," frequently referring to science fiction (Kelley et al., 2019). In the previously discussed 2018 survey of the US public, a majority of respondents assumed that Facebook photo tagging, Google Search, Netflix or Amazon recommendations, and Google Translate do not use AI (Zhang & Dafoe, 2019). In contrast, the majority of AI/ML researchers surveyed in Krafft et al. (2020) considered automated license plate readers and booking photo comparison software to use AI. The AI effect may eventually fade as existing applications of the technology, such as facial recognition software and large language models (e.g., OpenAI's GPT-3), become more salient in the news. Furthermore, criticism of predictive analytics that is not as advanced as AI, such as the UK grading algorithm and software used to make welfare eligibility decisions (Eubanks, 2018), are now incorporated in discussions around AI ethics and governance.

Trust in AI systems

The phrase "trustworthy AI" has become a ubiquitous phrase in AI ethics statements published by governments, tech companies, and civil society groups. Although the definition of trustworthy AI varies, common principles include beneficence, non-maleficence, autonomy, justice, and explicability (Thiebes et al., 2020). The buzz around trustworthy AI has produced various theoretical literature on designing algorithms and institutions that the public will trust. The chief shortcoming of these works is that they de-emphasize human users' subjective perceptions and experiences of those impacted by AI systems. At the same time, public opinion research has not kept up with these theoretical contributions; instead, they focus on the public's general attitude toward AI divorced of any technical or institutional context. Bridging the gap between these related areas of study would enhance our understanding of trustworthy AI.

Theoretical works on trustworthy AI frequently propose sociotechnical frameworks for building AI systems that the public will trust. These frameworks propose solutions that often involve building technical or institutional features to promote safety, fairness,

and transparency. Some proposed solutions include offering explanations for how the AI system works, creating documentation of the development process, requiring third-party audits, offering rewards for those who identify flaws in AI systems, creating a database of AI incidents, and enacting regulation (Brundage et al., 2020; Jacovi et al., 2021; Knowles & Richards, 2021). While these solutions could theoretically make AI systems safer and more ethical, their emphasis on the technical overlooks the role of subjective human judgment. In the classic "integrative model of organizational trust" (Mayer et al., 1995), ability, benevolence, and integrity are factors that increase *perceived* trustworthiness (emphasis mine). Furthermore, adding to the notion that trust is subjective, the authors argue that "[p]eople differ in their inherent propensity to trust" (p. 715). Consider the example of explaining how AI systems work: an experiment found that people with higher levels of education had a greater comprehension of the same explanations (Saha et al., 2020). Furthermore, even subject area experts were persuaded to trust AI systems when given misleading explanations (Lakkaraju & Bastani, 2020). Finally, individual differences in personality or technical expertise are correlated with different baselines level of trust in AI systems (Oksanen et al., 2020).

In contrast, much of the public opinion research has focused on respondents' subjective evaluation of AI in general. While these survey questions ask about respondents' perceived impact of AI on society or their support for developing AI, they could be measuring the general level of trust in AI. One central flaw in these studies is that they often ask about AI as a technology devoid of context, such as how or where the AI system is deployed. Nevertheless, they highlight differences in trust by geographic region, gender, and socioeconomic status.

Those living in East Asia, compared with other regions, have greater trust in AI across several comparative surveys. In a 2019 survey of over 150,000 respondents in 142 countries, 59 percent of those in East Asia indicated that AI would mostly help society, while 11 percent indicated that AI would mostly harm society. In contrast, in Latin America and the Caribbean, the region most wary of AI, 49 percent indicated that the technology would mostly help society while 26 percent indicated that it would mostly harm society (Neudert et al., 2020). These results have been replicated in another cross-national survey showing that those living in East Asian countries viewed the development of AI and workplace automation most positively (Johnson & Tyson, 2020). Content analysis of open-ended responses found that 14 percent of responses from South Korea described AI as "worrying," compared with 30 percent in the US and 31 percent in France (Kelley et al., 2019). In the US and the EU, where trust in AI systems is mixed, there is widespread consensus that AI is a technology that should be carefully managed (Zhang & Dafoe, 2020; Eurobarometer, 2017).

Two other important trends observed in these studies are that women and those of lower socioeconomic status (e.g., lower levels of education, lower levels of income) are less trusting of AI. In 15 out of 20 countries surveyed in Johnson and Tyson (2020), men, compared with women, indicated a significantly more positive view of AI development. In the same study, those with higher levels of education (having completed post-secondary education in high-income countries; having completed secondary education in middle-income countries) expressed a more positive view of AI development. In the US, those who earn more than $100,000 annually, compared with those in lower income brackets, expressed the highest level of support for developing AI at 59 percent. In contrast, only 33 percent of those earning less than $30,000 annually supported developing AI (Zhang & Dafoe, 2019).

Across 142 countries, business executives and other white-collared professionals, compared with those engaged in manual labor, were more likely to perceive AI as being helpful to society (Neudert et al., 2020). More research is needed to investigate how these differences in attitudes toward AI formed, and this will be discussed further in this chapter. One possible explanation is that AI systems have disproportionately harmed women, non-white people, and those of low socioeconomic status and excluded them from deciding how the technology is built and deployed (Gebru, 2020).

Given the novelty of AI as a topic for public opinion research, most cross-national studies ask about general attitudes toward AI devoid of technical or institutional context. Nevertheless, these studies reveal important variations in trust toward AI by country, gender, and socioeconomic status. Future survey work could examine how various subgroups in different countries perceive proposed strategies to make AI systems more trustworthy (e.g., by offering explanations or performing third-party audits). An additional advance in research would consider what drives perceived trust in AI systems deployed in different settings (e.g., facial recognition versus tagging abusive online content). While AI is often called a general-purpose technology, most AI systems deployed today have narrow applications. In the next section, I explore public opinion research that does not focus on AI in general but uses four specific applications: facial recognition, personalization algorithms, lethal autonomous weapons, and workplace automation.

Views toward Four Applications of AI

This section discusses the public opinion research on four highly salient applications in the news and has generated greater interest among survey researchers. Facial recognition technology and personalization algorithms have become commonplace but have lately generated much public scrutiny. Lethal autonomous weapons and advanced workplace automation are not yet ubiquitous but have received much attention in the press and popular media, given their potential to reshape international security and the economy. These four examples illustrate the need to study public opinion toward specific applications of AI, rather than AI as a general concept, as respondents often rely on their existing political heuristics when considering applications of AI in daily life.

Facial recognition

Facial recognition algorithms used to identify, verify, and classify persons based on their facial features have been deployed or approved for deployment in 109 countries as of 2020 (Surfshark, 2020). The technology has been standard in consumer applications, such as unlocking smartphones or tagging people in photos; now, law enforcement, employers, and businesses are increasingly turning to the technology as well. As facial recognition becomes more widespread, civil society groups and academic researchers have pointed out flaws in these AI systems and the risk to privacy and civil liberties. Researchers found that leading commercial facial recognition applications are much less accurate at identifying women,

particularly those with darker skin, than white men (Buolamwini & Gebru, 2018). Even if facial recognition algorithms were to become more accurate, critics contend that the technology would increase the capacity of law enforcement, governments, and even private companies to monitor the public—causing disproportionate harm to marginalized groups (West et al., 2018).

The public's view toward facial recognition technology is nuanced, although some key trends have been replicated across surveys. First, in several countries, the public is more supportive of facial recognition technology used by law enforcement compared with private actors. Second, support is correlated with demographic variables, such as the resident country, race, and political leaning.

Although much of the criticism of facial recognition technology has focused on its use and abuse by law enforcement, adults in several countries are more supportive of its use by law enforcement than businesses or employers. In the US, 59 percent of adults found uses by law enforcement to assess security threats in public spaces to be acceptable, while only 15 percent found uses by advertisers to track responses to ads acceptable (Smith, 2019). In the UK, 70 percent of adults supported uses in criminal investigations; in contrast, only seven percent supported uses by supermarkets to track shopper behavior, and four percent supported uses by employers to evaluate job candidates (Ada Lovelace Institute, 2019). Similarly, in Australia, more than 70 percent of adults supported uses by the police to monitor threats to society or to investigate crimes. In contrast, less than one-quarter of Australian adults supported uses by businesses to track customers or advertise to consumers (Automated Society Working Group, 2020). In a 2019 study of the public in China, Germany, the UK, and the US, support for central governments' use of facial recognition technology was higher than support for use by private businesses (Kostka et al., 2021).

Demographic variables, including country of residence, race, and political leaning, correlate with support for facial recognition technology. In the four-country survey discussed above, respondents in China indicated the highest support for facial recognition use (67 percent support). In contrast, 38 percent of adults in Germany, 50 percent in the UK, and 48 percent in the US supported the use of the technology (Kostka et al., 2021). Respondents in China, compared with the other countries, perceived facial recognition technology to be more convenient and efficient as well as less risky from privacy, discrimination, and surveillance perspectives. This finding is perhaps not surprising given the prevalence of the technology deployed by law enforcement and private actors in China.

Beyond these variations between countries, there are also differences in support among demographic subgroups within a country. For example, in a 2019 survey, while more than a majority of US adults supported law enforcement's use of facial recognition, support was much lower among Black respondents, younger generations, and those who identified with or leaned toward the Democratic Party (Smith, 2019). One explanation is that these demographic subgroups also have lower trust in law enforcement in general. Furthermore, US cities and states that have banned or placed a moratorium on the police's use of facial recognition technology are generally left-leaning in their politics (Recognition, 2020). As criticism of the technology becomes more politically salient, public opinion could change. For example, in the US, opposition to the use of facial recognition software has increased from 22 percent to 38 percent between 2018 and 2019 (Sabin, 2019).

Personalization algorithms

Personalization algorithms employ a user's online or offline data to create online content tailored to the user. Applications that use personalization algorithms include recommendation systems for news articles on social media, targeted online advertising, or individualized pricing. While personalization algorithms have become ubiquitous, researchers, journalists, and civil society groups have pointed out harms from the technology. From a privacy perspective, users' personal and behavioral data are collected without their informed consent by tech companies to sell to advertisers (Zuboff, 2019). Targeted advertising has excluded women and ethnic minorities from job listings and rental listings (Imana et al., 2021; Spinks, 2019). While concerns about "filter bubbles" on social media are overblown for the majority of users, personalization could be reinforcing the views of those with extremist political beliefs (Stark et al., 2020). This subsection reviews research that examines how users themselves understand and view personalization algorithms. Although public opinion research on this topic is growing, qualitative research has provided additional insights that address puzzles posed by survey data.

The works reviewed highlight the vast information asymmetry between the public and the tech companies that build and deploy personalization algorithms. The public lacks knowledge about the technology or even the vocabulary to talk about how these algorithms affect their online experiences. For example, nearly three-quarters of US Facebook users did not know that Facebook assigns them "interest categories" that are used to recommend them ads, news, and other content (Hitlin & Rainie, 2019). Furthermore, the majority of the US public did not perceive Facebook automated photo tagging or Netflix/Amazon's recommendation systems to involve AI or ML (Zhang & Dafoe, 2019). Qualitative interviews with 22 young people aged 16 to 26 revealed these frequent users of social media do not have a clear understanding of how personalization algorithms work (Swart, 2021). The researcher notes that while these young users lack the technical vocabulary to talk about personalization algorithms, the algorithms are objectively non-transparent. Qualitative interviews with "power users" (those who use a privacy/security browser extension to track and block data collection) revealed that even those with high levels of technical knowledge did not think they fully grasped the personalization algorithms they are resisting (Kant, 2020).

Although the public may not fully understand how personalization algorithms work, they oppose certain types of data from being collected and used or certain types of personalization. Data from nationally representative surveys in Germany, Great Britain, and the US found that the population in these countries opposed tech companies collecting sensitive information like personal tragedies or household incomes; in addition, the respondents expressed consensus against using personalization in political campaigning (Kozyreva et al., 2021). These survey results replicate findings from a survey of 748 Amazon Mechanical Turk workers that revealed respondents perceived the use of household income and race, compared with other types of data, in algorithmic personalization to be the most unfair (Coen et al., 2016).

The US public is not as opposed to the personalization of online newspaper front pages as those in Germany and Great Britain (Kozyreva et al., 2021). Nevertheless, 62 percent of US adults said that social media companies have too much control over the news people see, and 55 percent said these companies create a worse mix of news (Shearer & Grieco,

2019). Breaking down the data by party identification, Republicans expressed more negative views about social media platforms, with 85 percent indicating that these platforms censor their viewpoints—compared with 62 percent of Democrats (Smith, 2018). In reality, researchers do not find empirical evidence that platforms like Facebook, Twitter, or YouTube censor conservative viewpoints; in fact, these platforms go out of their way to appease conservatives in the US (Barrett & Sims, 2021).

Another gap between perception and reality is that many users want some personalization; yet, at the same time, they oppose tech companies collecting their data that are needed to build the personalization algorithms. Seventy-four percent of the public in Germany, along with 62 percent of the public in Great Britain and the US, exhibit this "acceptability gap" between personalized online services and data collection (Kozyreva et al., 2021). The researchers acknowledge that the public may not be aware that building personalization algorithms require collecting user data. Another view posited by Kant (2020) suggests that users are negotiating difficult trade-offs between protecting their privacy and accessing convenient online services; furthermore, users resign themselves to the fact that personalization algorithms are highly opaque and data collection is impossible to circumvent.

Lethal autonomous weapons

Lethal autonomous weapon systems can identify and engage targets without human intervention. The use of an autonomous drone (made by a Turkish company) in a March 2020 skirmish in Libya could possibly be the first deployment of this technology in battle (Vincent, 2021). Civil society groups, including the Campaign to Stop Killer Robots, have advocated for an international ban on fully autonomous weapons. These groups argue that lethal autonomous weapons are unethical and unsafe; furthermore, they suggest that an arms race to develop the technology would exacerbate tensions between major military powers. At the same time, the low cost of building lethal autonomous weapons could lead to proliferation among non-state actors, including terrorists (Warren & Hillas, 2020). Thirty countries have publicly expressed support for a pre-emptive international ban on fully autonomous lethal weapons. Still, several major military powers, including the US and Russia, currently oppose such a ban (Campaign to Stop Killer Robots, 2019).

Human Rights Watch and the Campaign to Stop Killer Robots have conducted two cross-national surveys that examine attitudes toward lethal autonomous weapons. While these organizations take an explicit policy position regarding the technology, their survey samples, compared with those in academic studies, contain the most diverse respondents in terms of geography. In 2018, 61 percent of those surveyed in 26 countries opposed the use of lethal autonomous weapons while 22 percent supported their use (Deeney, 2019). In a similar study conducted in 2017, 56 percent surveyed expressed opposition while 24 percent expressed support. Considerable cross-national variations exist: the 2018 survey found that support for fully autonomous weapons was highest in India (50 percent) and Israel (41 percent) and lowest in Turkey (12 percent) and the Netherlands (13 percent).

Academic studies find that US adults' support for lethal autonomous weapons can be affected by framing or new information. A 2013 survey experiment found that those who consume science fiction oppose lethal autonomous weapons when they are primed to

think about films that feature killer robots (Young & Carpenter, 2018). Survey experiments conducted in 2015 found that informing the US public that lethal autonomous weapons would be used to protect US troops increased support for developing the technology (Horowitz, 2016). The same paper reveals that informing the US public that foreign countries or non-state actors are developing these weapons also increased support. In contrast, informing US adults about the risks of an arms race between the US and China decreased respondents' support of US investing more in AI military technologies (Zhang & Dafoe, 2019). Future research could consider how other types of messaging would affect attitudes toward lethal autonomous weapons or expand the respondent pool of survey experiments to consider non-US publics.

Workplace automation

Concerns about workplace automation have existed throughout the 20th century but have recently intensified with the increasing focus on AI. According to the OCED, 14 percent of the jobs in 32 OECD countries are at high risk of being automated in the coming decades (Nedelkoska & Quintini, 2018). A more dire forecast puts the number at 47 percent of US jobs (Frey & Osborne, 2017). Survey data over time show that the US public's views toward workplace automation have become more uncertain in recent years. The US National Science Foundation (NSF) had conducted eight surveys between 1983 and 2003, asking respondents whether they agree or disagree that computers and factory automation will create more jobs than they will eliminate. Survey data from Zhang and Dafoe (2019) showed similar levels of disagreement with the NSF survey statement (around one-half of the respondents disagreed) but a higher percentage of respondents who indicated they do not know (24 percent in 2018 versus less than 10 percent in all of the NSF surveys).

Recent survey research has attempted to disentangle the fear of automation in general versus the fear of one's own job becoming automated. Workers in the US expressed optimism bias regarding automation: they believe that while many jobs are likely to be automated, their own will be safe from automation (Smith & Anderson, 2016). Respondents whose jobs are objectively more likely to be automated do not think that their jobs are at higher risk of automation. The correlation between workers' forecasts and some economic forecasts is as low as 0.11 (Zhang, 2022). Furthermore, workers indicated they are more worried about losing their jobs to cheaper labor than being replaced by computers and machines (Smith, 2016).

One set of findings among recent observational studies is that actual or anticipated exposure to automation is positively correlated with support for right-wing populist parties, candidates, or policies. Comparative analysis using regional-level and individual-level data from several European countries found that those more exposed to automation shocks indicated greater support for nationalist and radical-right parties (Anelli et al., 2019; Im et al., 2019), even after accounting for social welfare programs that potentially helped workers harmed by automation (Gingrich, 2019). In the US, exposure to industrial robots is positively correlated with support for Donald Trump in the 2016 Presidential Election at the electoral district level (Frey et al., 2018). Individual-level survey data suggests that Americans who are more exposed to automation expressed more significant opposition to free trade and immigration (Wu, 2021).

The link between fear of automation and right-wing politics is not replicated across all studies. Other studies found that workers exposed to automation risks are more in favor of left-wing policies or parties—yet other papers found that the threat of automation does not shift political preferences at all. An analysis of survey data from 17 European countries between 2002 and 2012 finds that respondents whose jobs were more automatable expressed greater support for redistribution (Thewissen & Rueda, 2019). Exposure to automation is positively correlated with support for not only radical right-wing parties but also mainstream left-wing parties (Gingrich, 2019). In a survey experiment, an informational treatment that made American respondents more aware of automation's threat *increased* support of universal basic income (UBI) among low-skilled workers (Lekalake et al., 2019). Other studies find null effects. In another survey experiment, exposing US workers to news articles about how automation will threaten jobs in general and their individual jobs did not shift support for expanding the welfare state or UBI (Zhang, 2022). In an observational study of 21 European countries, researchers found no association between risk of job automation and UBI support (Dermont & Weisstanner, 2020).

Given workers' uncertainty about how AI will impact their jobs, it seems reasonable that this set of nascent research papers would reach different conclusions. A shortcoming of these research papers is that they focus on OECD countries while ignoring workers in the Global South. Future survey research should consider expanding the geographic scope by surveying workers in middle and low-income countries.

Directions for Future Research

This chapter attempts to synthesize the existing research on the public's attitudes toward AI. An increasing number of cross-national studies allow researchers to explore variations in attitudes between countries and demographic subgroups. Research on knowledge about AI revealed that the public is increasingly aware of AI but tends to anthropomorphize the technology. In terms of general attitudes toward AI, those living in East Asian countries, men, and those of higher socioeconomic status tend to have more positive views toward AI. Researchers have also branched out to study specific AI applications, such as facial recognition technology, personalization algorithms, lethal autonomous weapons, and workplace automation. This latter set of research is crucial for revealing that attitudes are shaped by the public's existing beliefs about politics and ethics; furthermore, the public's attitudes about AI can be affected by messaging.

Nevertheless, there is much potential for future research that expands upon existing studies. In this section, I propose four new directions that seek to answer some fundamental unanswered questions regarding public attitudes toward AI.

Institutional trust in actors behind AI systems

First, researchers could explore institutional trust in actors behind AI systems within the contemporary political and economic context. This line of research somewhat differs from empirically testing the theories of trustworthy AI, which tend to generate abstract solutions

and de-emphasize the power struggle between tech companies, governments, and the public. Given current policy debates around regulating major tech companies and technological competition between the US and China, empirical studies must not be agnostic to the actors in this space. In fact, survey research in the US shows that the public has different levels of trust in actors to build AI systems (Zhang & Dafoe, 2019). US adults place the greatest amount of trust in university researchers and the military to build AI; furthermore, they place greater trust in tech companies than the government. However, trust in tech companies is not uniform: the public places significantly less trust in Facebook than other major tech companies.

Future research could try to explain such variations in trust in actors building AI systems. Recently, there has been increasing backlash against major tech companies over their disproportionate market dominance as well as their failure to protect user data and prevent the spread of dis/misinformation. Research questions could examine whether the overall reputation of a tech company affects the public's perception of its AI products. For instance, would the public place less trust in AI systems developed by a company that has repeatedly been criticized for having data breaches or problematic content moderation practices?

Impact of experience and knowledge on attitudes

The second research direction is to examine how increasing users' experience with and knowledge about AI will impact their attitudes toward the technology. Many governments and civil society groups have proposed educating the public about AI to empower citizens. Furthermore, various national AI strategies have called for educating students about AI to train a competent workforce for the future. Finally, as more AI systems become deployed in the real world, the public will increasingly interact with AI applications online, in public, or at their workplaces. This review has shown that making generalizations about how increased experience and knowledge impact attitudes toward AI is difficult. Further theoretical and empirical work should take a more nuanced approach.

For instance, the relationship between technical knowledge about AI and trust in AI systems does not appear to be monotonic increasing. US adults without technical training tend to be less supportive of AI and more concerned about AI governance challenges compared with US adults who have a computer science or engineering degree (Zhang & Dafoe, 2019). Yet AI/ML researchers are increasingly aware of the dangers and harmful societal consequences of AI systems (Belfield, 2020); furthermore, an international sample of AI researchers—compared with the US public—are less trustful of tech companies to develop and manage AI (Zhang et al., 2021). Future research could test whether knowledge about AI and trust in AI systems follows an inverted-U shape: increasing knowledge increases trust up to a point then decreases as one becomes an expert.

Another aspect of this research direction is to examine how different types of experience or education impact attitudes toward AI. The impact of experience on trust varied by the type of AI the user interacts with, according to a systemic literature review (Glikson & Woolley, 2020). For virtual AI and embedded AI (AI that is embedded in sociotechnical systems and not visually visible to users), their trust started high and decreased with use. In contrast, for robotic AI, their trust started low and increased with use. The authors of the review paper acknowledged these trends are typically observed in short-term studies

and proposed that future research track how experience impact trust with long-term use. Research has also shown that using an AI application did not necessarily increase knowledge of how AI systems work. For example, those who frequently use social media did not understand how platforms use their data to generate personalized content or to categorize users (Hitlin & Rainie, 2019; Swart, 2021). Therefore, trust might be mediated through subjective user experience and not necessarily a technical assessment of the AI's capability or safety.

Finally, when educating people about AI, the type of information conveyed (e.g., technical knowledge versus information about the societal impact of AI) could affect attitudes differently. Papers presented at ML conferences tend to focus on improving the performance, generalization, and efficiencies of models (Birhane et al., 2021), rather than making AI systems safer, fairer, or more explainable—topics that are prevalent at AI ethics conferences. Future research could test how taking a course on AI ethics or attending an AI ethics conference (versus taking a computer science class on AI or attending an ML conference) would impact students' trust in AI and their views toward AI ethics.

Explaining variations in attitudes

Beyond knowledge and expertise, I have observed other correlates of attitudes toward AI. As noted previously, women and those of lower socioeconomic status tend to be less supportive of AI. As also discussed, partisanship and race can affect attitudes toward specific applications of AI, such as facial recognition technology. These studies that identify how different demographic subgroups view AI show that the public is not monolithic in their attitudes. But more research is needed to understand the mechanisms that cause different demographic subgroups to view AI differently.

One potential mechanism is self-interest, or rather self-preservation. As previously mentioned, one explanation for women's relatively low level of trust in AI is that researchers and civil society groups have found that numerous AI systems are discriminatory toward women. Facial recognition systems are less accurate for women, particularly women with darker skin (Buolamwini & Gebru, 2018). Furthermore, AI systems used to target job advertisements and evaluate job applications have excluded women from being hired (West et al., 2018). These and other examples could explain why women question that AI will be beneficial to them. Besides material self-interest explanations, the gender difference in trust in AI could be explained by gender differences toward science and technology in general. In a global survey covering 20 countries, men, compared with women, tend to be more supportive of nuclear power and perceive GM foods to be safer (Funk et al., 2020). Finally, attitudes toward AI could also be driven by factors not fully conscious to respondents, such as personality traits. A large-*N* online lab experiment found that of the Big Five personality traits, openness to experience is positively correlated with trust in robots, and conscientiousness is negatively correlated with trust (Oksanen et al., 2020). Future research could investigate other domains of personality beyond the Big Five, such as authoritarian personality, religiosity, and risk-seeking.

Instead of explaining differences in attitudes toward AI by demographic subgroups, we could use cluster analysis to identify groups within the public who share similar views toward AI. Unlike subgroup analysis, cluster analysis involves unsupervised classification

where researchers do not pre-define groups or even the number of groups. Cluster analysis has been used to segment the public into six groups based on their views toward climate change (Maibach et al., 2009; Chryst et al., 2018). Similar research using survey data about AI could help researchers move beyond the techno-optimist versus techno-pessimist binary that is an oversimplification of AI policy discourse. Cluster analysis could help us identify groups in the population who view AI as beneficial but prefer greater regulation to protect consumers and citizens. Furthermore, such research would allow one to estimate the percentage of the population who belongs to each cluster and what demographic subgroups make up the clusters.

Impact of beliefs on consumer and civic behavior

The fourth new research direction for future research is to consider how beliefs about AI impact consumer and civic behavior. More and more AI systems are embedded within products or services that the public can purchase or deployed in public spaces. Researchers studying human trust in AI have used short-term, small-sample experiments to test whether explaining how the AI makes decisions will increase trust (Glikson & Woolley, 2020). Deploying these experiments in real-world settings and taking multiple measures over time could help researchers understand how stated preferences relate to behavior. For instance, would informing consumers about the risks and benefits of using an AI-powered mental health chatbot affect people's willingness to use it? Would educating people about racial/gender bias in facial recognition technology increase the likelihood of their signing a petition to ban its use by law enforcement?

This line of inquiry could draw upon human–computer interaction research on privacy and consumer behavior. One central finding in this literature is the privacy paradox: although consumers indicate that they care about privacy, they take little effort to protect their privacy online (Barth & De Jong, 2017). More recent work pushes back against the idea that consumers are irrational by arguing that privacy policy statements are too complex for consumers to understand (Bashir et al., 2015) and that consumers have grown too cynical about tech companies' willingness or ability to protect their data (Hoffmann et al., 2016). We could observe a similar "AI ethics paradox" in future studies where the public expresses deep concerns about AI systems causing harm yet fail to take action as consumers or citizens. If anything, the information asymmetry between developers and the public is even more significant in the AI realm than in the privacy realm: black-box algorithms are even more incomprehensible to laypeople than lengthy privacy policies.

The four new research directions discussed above are just some ways that scholars of public opinion can advance research on this topic. Researchers can draw upon the ever-expanding literature on AI ethics and governance from ML, information science, and science and technology studies to develop new surveys and experiments.

Acknowledgments

I am grateful for feedback from Laurin Weissinger, Toby Shevlane, Nataliya Nedzhvetskaya, Markus Anderljung, Noemi Dreksler, Justin Bullock, and the CIFAR Innovation, Equity & the

Future of Prosperity Program (particularly Ray Gosine, Goldie Nejat, Dan Breznitz, Amos Zehavi, David Y. Yang, J. Bradford DeLong, Daniel Aldana Cohen, and Kenneth Lipartito). I would also like to thank Aura Gonzalez for her helpful research assistance. Collaborative research with Allan Dafoe, Michael C. Horowitz, Markus Anderljung, Noemi Dreksler, and Lauren Kahn has inspired this work. My research has been supported by the CIFAR Azrieli Global Scholars Program and the Ethics and Governance of AI Fund.

References

Ada Lovelace Institute. (2019). Beyond face value: Public attitudes to facial recognition technology (Tech. Rep.). Ada Lovelace Institute. https://perma.cc/M4EG-N5MY.

Anelli, M., Colantone, I., & Stanig, P. (2019). We were the robots: Automation in manufacturing and voting behavior in western Europe. https://www.iza.org/publications/dp/12485/we-were-the-robots-automation-and-voting-behavior-in-western-europe (Working Paper).

Automated Society Working Group. (2020). Australian attitudes to facial recognition: A national survey (Tech. Rep.). Monash University. https://perma.cc/BE3C-T9W9.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature 563*(7729), 59–64.

Barrett, P. M., & Sims, J. G. (2021). False accusation: The unfounded claim that social media companies censor conservatives (Tech. Rep.). New York University, Stern School of Business, Center for Business and Human Rights. https://perma.cc/X5A8-HQLB.

Barth, S., & De Jong, M. D. (2017). The privacy paradox—investigating discrepancies between expressed privacy concerns and actual online behavior—a systematic literature review. *Telematics and Informatics 34*(7), 1038–1058.

Bashir, M., Hayes, C., Lambert, A. D., & Kesan, J. P. (2015). Online privacy and informed consent: The dilemma of information asymmetry. *Proceedings of the Association for Information Science and Technology 52*(1), 1–10.

BBC. (2017). *Switzerland votes to phase out nuclear power*. https://perma.cc/52TE-QHNG.

Belfield, H. (2020). Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15–21). New York, NY, USA: Association for Computing Machinery.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). The values encoded in machine learning research. arXiv preprint arXiv:2106.15590.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T. . . . Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91), New York, NY, USA.

Campaign to Stop Killer Robots. (2019). *Country views on killer robots*. Author. https://perma.cc/S4MD-P4BZ.

Cave, S., Coughlan, K., & Dihal, K. (2019). Scary robots: Examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 331–337). New York, NY, USA: Association for Computing Machinery.

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). Portrayals and perceptions of AI and why they matter (Tech. Rep.). The Royal Society. https://perma.cc/7NMY-PBYX.

Chryst, B., Marlon, J., van der Linden, S., Leiserowitz, A., Maibach, E., & Roser-Renouf, C. (2018). Global warming's "Six Americas Short Survey": Audience segmentation of climate change views using a four-question instrument. *Environmental Communication 12*(8), 1109–1122.

Coen, R., Paul, E., Vanegas, P., Lange, A., & Hans, G. (2016). A user-centered perspective on algorithmic personalization (Tech. Rep.). https://perma.cc/AE2Q-PRWB.

Cooper, A. F., & Abrams, E. (2021). Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 46–54). New York, NY, USA: Association for Computing Machinery

Deeney, C. (2019). Six in ten (61%) respondents across 26 countries oppose the use of lethal autonomous weapons systems. https://perma.cc/6N63-EUHN.

de Figueiredo, A., Simas, C., Karafillakis, E., Paterson, P., & Larson, H. J. (2020). Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: A large-scale retrospective temporal modelling study. *The Lancet 396*(10255), 898–908.

Dermont, C., & Weisstanner, D. (2020). Automation and the future of the welfare state: Basic income as a response to technological change? *Political Research Exchange 2*(1), 1757387.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Eurobarometer. (2017). Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life (Tech. Rep.). https://perma.cc/9FRT-ADST.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI (Tech. Rep.). Berkman Klein Center for Internet & Society. https://ssrn.com/abstract=3518482orhttp://dx.doi.org/10.2139/ssrn.3518482.

Frewer, L., Lassen, J., Kettlitz, B., Scholderer, J., Beekman, V., & Berdal, K. G. (2004). Societal aspects of genetically modified foods. *Food and Chemical Toxicology 42*(7), 1181–1193.

Frey, C. B., Berger, T., & Chen, C. (2018). Political machinery: Did robots swing the 2016 us presidential election? *Oxford Review of Economic Policy 34*, 418–442.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114*, 254–280.

Funk, C., Tyson, A., Kennedy, B., & Johnson, C. (2020). Science and scientists held in high esteem across global publics (Tech. Rep.). Pew Research Center. https://perma.cc/JYE7-N7ZX.

Gebru, T. (2020). Race and gender. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of AI ethics* (pp. 253–269). Oxford University Press.

Gingrich, J. (2019). Did state responses to automation matter for voters? *Research & Politics 6*(1), 2053168019832745.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals 14*(2), 627–660.

Guariglia, M. (2020). What to know before you buy or install your Amazon Ring Camera. Electronic Frontier Foundation. https://perma.cc/4M87-VFUB.

Hitlin, P., & Rainie, L. (2019). Facebook algorithms and personal data (Tech. Rep.). Pew Research Center. https://perma.cc/S7YG-PPA5.

Hoffmann, C. P., Lutz, C., & Ranzini, G. (2016). Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace 10*(4), 43–60.

Horowitz, M. C. (2016). Public opinion and the politics of the killer robots debate. *Research & Politics 3*(1). https://doi.org/10.1177/2053168015627183.

Im, Z. J., Mayer, N., Palier, B., & Rovny, J. (2019). The "losers of automation": A reservoir of votes for the radical right? *Research & Politics 6*(1), 2053168018822395.

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021* (pp. 3767–3778).

Jacovi, A., Marasovi´c, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635).

Johnson, C., & Tyson, A. (2020, December). People globally offer mixed views of the impact of artificial intelligence, job automation on society (Tech. Rep.). Pew Research Center. https://perma.cc/6NYT-DH4M.

Kant, T. (2020). *Making it personal: Algorithmic personalization, identity, and everyday life*. Oxford University Press.

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T., & Woodruff, A. (2019). "Happy and assured that life will be easy 10years from now.": Perceptions of artificial intelligence in 8 countries. arXiv preprint arXiv:2001.00081.

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262–271).

Kolkman, D. (2020). F**k the algorithm?: What the world can learn from the UK's A-level grading fiasco. London School of Economics and Political Science. https://perma.cc/CLL6-QE6N.

Kostka, G., Steinacker, L., & Meckel, M. (2021). Between security and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States. *Public Understanding of Science*, *30*(6), 671–690.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications 8*(1), 1–11.

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 72–78). New York, NY, USA: Association for Computing Machinery.

Kulynych, B., Madras, D., Milli, S., Raji, I. D., Zhou, A., & Zemel, R. (2020, July). Participatory approaches to machine learning. In *International Conference on Machine Learning Workshop*. https://participatoryml.github.io/.

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79–85). New York, NY, USA: Association for Computing Machinery.

Lekalake, R., Markovich, Z., Nahmias, G., & Russell, S. (2019). Automation risk and support for a universal basic income (Working Paper).

Maibach, E., Roser-Renouf, C., & Leiserowitz, A. (2009). Global warming's six Americas 2009: An audience segmentation analysis (Tech. Rep.). Yale Project on Climate Change and Center for Climate Change Communication, George Mason University. https://perma.cc/C6DL-HE3Y.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review 20*(3), 709–734.

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

Moody, B. (2011). Italians say no to nuclear energy in referendum. *Reuters*. https://perma.cc/P89B-RFPV.

Nedelkoska, L., & Quintini, G. (2018). Automation, skills use and training (Tech. Rep.). OECD iLibrary. https://doi.org/10.1787/2e2f4eea-en.

Neudert, L.-M., Knuutila, A., & Howard, P. (2020). Global attitudes towards AI, machine learning & automated decision making (Tech. Rep.). Oxford Internet Institute. https://perma.cc/6PB6-X56B.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology 11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256.

Olive, J. K., Hotez, P. J., Damania, A., & Nolan, M. S. (2018). The state of the antivaccine movement in the United States: A focused examination of nonmedical exemptions in states and counties. *PLoS Medicine 15*(6), e1002578.

Pelinka, A. (1983). The nuclear power referendum in Austria. *Electoral Studies 2*(3), 253–261.

Recognition, B. F. (2020). Ban facial recognition map. https://www.banfacialrecognition.com/map/.

Requejo, J., Griffiths, U., Duncan, R., Mirza, I., & Mebrahtu, S. (2020). Immunization coverage: Are we losing ground? (Tech. Rep.). UNICEF. https://perma.cc/6NL8-YXJW.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach*. Prentice Hall.

Sabin, S. (2019). Voter support for facial recognition technology slips amid debates over its use. Morning Consult. https://perma.cc/LMQ5-XXTY.

Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L., & Tschantz, M. C. (2020). Human comprehension of fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 152–152). New York, NY, USA: Association for Computing Machinery.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience 11*(2), 88–95.

Shearer, E., & Grieco, E. (2019). Americans are wary of the role social media sites play in delivering the news (Tech. Rep.). Pew Research Center. https://perma.cc/JLH2-C6H4.

Smith, A. (2016). Public predictions for the future of workforce automation (Tech. Rep.). Pew Research Center. https://perma.cc/XN9V-M4L2.

Smith, A. (2018). Public attitudes toward technology companies (Tech. Rep.). Pew Research Center. https://perma.cc/96GP-7VFS.

Smith, A. (2019). More than half of U.S. adults trust law enforcement to use facial recognition responsibly (Tech. Rep.). Pew Research Center. https://perma.cc/FUV7-5BDJ.

Smith, A., & Anderson, M. (2016). Automation in everyday life (Tech. Rep.). Pew Research Center. https://perma.cc/WU6B-63PZ.

Spinks, C. N. (2019). Contemporary housing discrimination: Facebook, targeted advertising, and the fair housing act. *Houston Law Review 57*, 925.

Stark, B., Stegmann, D., Magin, M., & Ju¨rgens, P. (2020). Are algorithms a threat to democracy? the rise of intermediaries: A challenge for public discourse (Tech. Rep.). AlgorithmWatch. https://perma.cc/WS6G-NQ5U.

Surfshark. (2020). The facial recognition world map (Tech. Rep.). Surfshark. https://perma.cc/6HUK-9Q4D.

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media + Society 7*(2), 20563051211008828.

Thewissen, S., & Rueda, D. (2019). Automation and the welfare state: Technological change as a determinant of redistribution preferences. *Comparative Political Studies 52*(2), 171–208.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*, 447–464. https://doi.org/10.1007/s12525-020-00441-4.

Vincent, J. (2021, June). Have autonomous robots started killing in war? *The Verge*. https://perma.cc/K4CU-PWL3.

Warren, A., & Hillas, A. (2020). Friend or frenemy? the role of trust in human-machine teaming and lethal autonomous weapons systems. *Small Wars & Insurgencies 31*(4), 822–850.

West, S. M., Whittaker, M., & Crawford, K. (2018). Discriminating systems: Gender, race, and power in AI (Tech. Rep.). AI New Institute. https://perma.cc/3VAJ-E38X.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200). New York, NY, USA: Association for Computing Machinery.

Wu, N. (2021). Misattributed blame? Attitudes toward globalization in the age of automation. *Political Science Research and Methods*, 1–18. doi:10.1017/psrm.2021.43.

Young, K. L., & Carpenter, C. (2018). Does science fiction affect political fact? yes and no: A survey experiment on "killer robots." *International Studies Quarterly 62*(3), 562–576.

Zhang, B. (2022). No rage against the machines: Threat of automation does not change policy preferences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 856–866). Oxford, UK: Association for Computing Machinery.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research 71*, 591–666.

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends (Tech. Rep.). Centre for the Governance of AI, University of Oxford. http://dx.doi.org/10.2139/ssrn.3312874.

Zhang, B., & Dafoe, A. (2020). US public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 187–193). New York, NY, USA: Association for Computing Machinery.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

CHAPTER 29

ADDING COMPLEXITY TO ADVANCE AI ORGANIZATIONAL GOVERNANCE MODELS

JASMINE MCNEALY

Introduction

In May 2021, United States Senator Markey and Representative Matsui proposed legislation aimed at banning discriminatory algorithmic processes on online platforms (Dhar, 2021; Zakrzewski, 2021). Called the *Algorithmic Justice and Online Platform Transparency Act*, the bill, if passed, would have banned the use of algorithmic processes that discriminate on the basis of race, age, gender, ability, and other protected characteristics, as well as the establishment of standards for safety and effectiveness for algorithms. Senator Markey has said that the aim of the bill was "to open up Big Tech's hood, enact strict prohibitions on harmful algorithms, and prioritize justice for communities who have long been discriminated against as we work toward platform accountability" (*Senator Markey, Rep. Matsui Introduce Legislation to Combat Harmful Algorithms and Create New Online Transparency Regime*, 2021). Representative Matsui, too, took a justice-oriented perspective when detailing the need for the law, saying, "We stand amidst a reckoning on racial justice and discrimination, and we must seize the moment by doing all we can to root out prejudiced practices wherever they occur. As we work for justice and reform, it is crucial that we remain vigilant by demanding transparency from 21st century platforms about the algorithms that shape our online interactions."

Industry professionals have taken a more cautious tone when speaking about how the law would work in practice. One of the requirements of the law would force platform transparency—notice and explanation of how personal data collected on the site would be used and the algorithmic processes involved with use. Critics of laws forcing transparency argue that although transparency or making users aware of the data collection and the use process is an admirable goal, algorithmic processes are complex. This means that even with

information available, the average person may not be able to understand what happens within the processes. Instead, commenters have called for making the usual black box artificial intelligence information available to oversight organizations including advocacy organizations, universities, and governments.

For its part, the *AJOPT* Act would create an inter-agency task force composed of entities including the Department of Education, Department of Housing and Urban Development, Department of Commerce, and others, tasked with oversight and investigation of discriminatory algorithms. At the same time, state attorneys general would be able to continue to exercise their powers to protect consumers in their individual states. The creation of this kind of task force and allowance for state rights enforcement is representative of the possibilities of organizational governance models that are a mixture of several different agencies with the aim of more effectively governing a complex subject area. This chapter considers the need for more complexity as a possible solution for the difficulty of governing artificial intelligence. First, this chapter examines different kinds of traditional organizational governance, examining the benefits and critiques by investigating examples of each and its implementation. Following this is an overview of an example of three attempts at the creation of AI regulation that include some kind of organizational governance scheme. Next, this chapter discusses the opportunities for complexity in AI governance and examples of policy and governance schemes being proposed.

Governance, As A General Matter

When considering the best model of governance for use with emerging technology, particularly artificial intelligence, it is important to consider what governance entails. Governance, within the context of this chapter, refers to statecraft, which Stiver (2008, p. 5) defines as "the art of acting according to duty, justice, and reason on behalf of a community of citizens." This means that policy and regulatory creations are focused on reaching higher goals like "truth, fairness, and democracy" (Robichau, 2011, p. 115). Reaching these goals requires encouragement and protection of civic engagement (Stiver, 2008), as well as creativity and accountability.

Traditional governance models

A demonstration of statecraft can be seen, then, in the movement and successes of grassroots, community, and advocacy organizations in convincing some local governments to ban the use of facial recognition technology. Between 2018 and 2021, several U.S. municipalities, including Oakland, San Francisco, Boston, and Jackson, Mississippi banned the use of facial recognition technologies. These prohibitions arose from the work of coalitions of community groups, concerned citizens, and advocacy organizations petitioning city council members to take up the matter of the use of facial recognition technology by law enforcement and other government entities, and the unfairness that may result. Many of the cities with bans are more populace or are suburban areas connected to larger metropolitan areas.

At the same time, these cities are home to diverse populations, which makes the risk of erroneous use of the AI technology even more probable.

The work by concerned citizens and activists to have municipal bans on law enforcement's use of facial recognition technology demonstrates the creation of policy in accord with ideas of justice, especially on behalf of traditionally marginalized or vulnerable populations. But although these bans were successful, the methods of controlling technology are traditional, reflective of a common governance scheme: command-and-control. Used popularly in connection with environmental regulations, the command-and-control approach is a frame in which regulators make commands in the form of law and regulation, and then control how these commands are achieved (Cole & Grossman, 2018). These controls usually require the use of technology. In the environmental context, this would include placing sensors on exhaust stacks to collect pollution data or using sensors and cameras on fishing boats to ensure the correct fish counts are reported.

Command-and-control governance is hierarchical. Commands, in the form of law, regulation, policy, guidelines, etc. come from a central place (Cox, 2016). This central place has not only the concentration of power but also the ability to enforce the regulations that it promulgates. The specter of enforcement through civil or criminal penalties motivates organizations and individuals to comply with regulations (Li et al., 2010). But top-down governance schemes have significant drawbacks. The primary criticism of command-and-control approaches to governance is that they are inefficient and costly (Bevir, 2012; Sinclair, 1997). Further, the scheme promotes regulations that are often inflexible and become obsolete in a time of rapid innovation. This is particularly so when considering data governance, where advancements in how data is collected, accessed, and otherwise changes at light speed.

A second common model is that of market governance, in which "decisions are made within the framework of overarching policies" (Pierre & Peters, 2019, p. 1). This means that, unlike command-and-control and other hierarchical systems, power is not concentrated at the top, but that there are structures in place to enforce policy. Market governance has also been defined as an environment in which "economic actors can cooperate to resolve common problems without distorting the basic mechanisms of the market" (Pierre & Peters, 2019, p. 10). In market governance schemes, organizations are incentivized to behave in certain ways through both formal and informal rules. This can create a system that is more flexible than the traditional command-and-control model of governance.

Although any market governance incentives or mechanisms include instruments like taxes, subsidies, and refund systems, etc., the schemes are not devoid of regulation. Regulations are, however, meant to promote and motivate compliance. Regulation is one of the three roles Gereffi and Mayer (2006) identified as important in market governance along with facilitation and compensation. Regulations place boundaries on organizations and organizational schemes that could cause harm or have other negative implications. Facilitation assists with the establishment of organizational rights like property rights, as well as rules of competition and contract enforcement, among other things. Compensation focuses on mediating disparities in market impacts. This would include the creation of welfare policies like healthcare, social insurance, and public education. It is important to note also that market governance can be both public and private and/or have public and private components. Public market governance is demonstrated through law, regulation, and the

implementation of programming; private governance through voluntary codes, standards, and philanthropy.

Advanced governance

A third kind of governance model is that of networked governance. Whereas command-and-control uses a top-down approach, networked governance is more bottom-up, requiring groups of organizations and individuals to work in concert toward a goal. Centering collective decision-making that proposes to include different kinds of participants, network governance works to govern complex and uncertain systems (Stoker, 2006). Governance can begin as a creation of the state or civil society organizations. All kinds of networks include a participatory frame, engaging various stakeholders. This kind of governance has been cited as advantageous for capacity building, better services, and addressing complex problems, among other things (Alter & Hage, 1993; Brass et al., 2004).

Multiple forms of governance networks exist (Sørensen & Torfing, 2009). Provan and Kenis (2008) identified three kinds of governance networks: participant-governed, organization-governed, and administratively governed. Participant-governed networks can be governed by collective member organizations or by one organization that takes the lead, through formal meetings or more informally. These networks depend on member engagement. Importantly, power in decision-making is symmetrical among all member organizations (Provan & Kenis, 2008). Networks governed by lead organizations appear more like a centralized governance model. Yet, the lead organization is only tasked with coordinating decision-making, not making the decisions for the network. That said, power is asymmetrical. Finally, administratively governed networks are those that set up an external, separate administrative organization, not a member organization, to broker the network.

Whatever the form of network governance Sorensen and Torfing (2016) identified five important elements of a governance network:

> (1) A relatively stable horizontal articulation of interdependent, but operationally autonomous actors; (2) who interact through negotiations that involve bargaining, deliberation and intense power struggles; (3) which take place within a relatively institutionalized framework of contingently articulated rules, norms, knowledge and social imaginaries; (4) that is self-regulating within limits set by external agencies and; (5) which contribute to the production of public purpose in the broad sense of visions, ideas, plans and regulations." (2016, p. 9)

Hybridity in network governance means that the scheme benefits from the positives of each model, while at the same time suffering from any of the frailties. The mixture of models is supposed to help remedy many of the drawbacks of the separate schemes. Often this means that non-state actors take on roles traditionally filled by state agencies or engage with state actors in ways that make it difficult to separate state from non-state. According to Goodfellow and Lindemann (2013), hybridity is not the same as multiplicity, or many actors coexisting and/or in competition in one space. Instead, it is the possibility of multiple actors working together with their own logics to create new logics.

An example of a kind of hybrid model, and an arguably ineffective governance scheme, is the regulation of privacy and data protection in the United States. Currently, no omnibus federal privacy legislation exists, which means a mix of state and federal agencies and state

attorneys general, as well as state and federal legislation and regulation attempt to remedy possible harms from data misuse, collection, and storage. Although no omnibus legislation exists mandating a specific federal agency take the lead on privacy and data protection enforcement, some agencies have required business transparency. At the same time, and under the same privacy threats, many federal agencies alone are not equipped or funded to take on all data practices that evoke individual privacy. Also, individuals have no private right of action for many of the federal regulations in existence that touch privacy related issues. This does not mean that individuals are left completely without recourse; state and other laws may allow civil suits or state attorneys general to pursue criminal penalties. This does demonstrate that hybrid systems can suffer from failures familiar to traditional forms of governance.

In comparison, hybrid governance models have been deemed successful in research on civic technology resources including cybersecurity governance (Pylant, 2020). Prior literature had indicated that a centralized governance approach for cybersecurity would be the most effective for state governments. Pylant (2020) found, however, that cybersecurity centralization is constrained by budget shortfalls, cultural issues, and lack of personnel, all of which make centralization unfeasible. Using a comparisons of state scores from the Nationwide Cybersecurity Review (NCSR), where states with higher scores were identified as having more successful cybersecurity programs), the study suggested that hybrid governance could also make for successful state cybersecurity programs. Hybridity, in this study, meant that instead of one central agency, there was a mix between a central agency and sub-agencies with cybersecurity authority. In certain instances—state-specific contexts—the study found that a hybrid governance model could be more effective than centralized authority.

Although the definitions of hybrid, networked governance until now have focused on the interrelatedness of the organizations that would oversee enforcement, governance strategies, also, can reflect the networked nature of a problem. That is, the terms of the regulation can and should be such that the law is not faced with an implementation gap—the divide between policy goals and how it works in practice (Bayrakal, 2006; Braithwaite et al., 2018; Hudson et al., 2019; McNealy et al., 2022). Networked governance, in this sense, includes policies that appreciate the complexity of the issue of focus. Braithwaite et al. (2018) argued that complexity, like in health care, requires a different kind of perspective on implementing changes. While interventions, like policy creation, make up a portion of the process of modification, without an adequate understanding of the components—people, technology, artifacts, equipment—in a system, it will be difficult to make adequate change.

AI and Governance

Traditional forms of governance, likewise, may not withstand the complexity of regulating artificial intelligence and its associated technologies. AI is a complex and multifaceted technology that appears in various forms and across myriad industries and sectors. It is both facial recognition technology and movie recommendations, autonomous vehicles, and virtual reality dressing rooms. If this were not enough to evoke the need for governance models able to deal with the many faces of AI, then add the additional layers of multiple governments,

sectors, stakeholders, and not the least, those impacted by AI uses. Complexity, specifically findings from studies in complexity science, might offer a way of rethinking legislative proposals for governance organization schemes that are better equipped for emerging technology. Currently, governments at the state, country, and regional levels are proposing or announcing policies related to AI and algorithmic technologies. This section briefly describes three examples of recently proposed AI regulation: the Artificial Intelligence Act from the EU, the Brazilian Artificial Intelligence Bill, and the U.S. Algorithmic Accountability Bill.

Europe: Artificial Intelligence Act

The European Commission's draft Artificial Intelligence Act provides an important example for consideration (AI Act, 2021). Proposed in April 2021, the AIA is supposed to be the European Union's answer to technology that has become pervasive and is impacting several different areas of society, especially in social services. The proposed law offers a risk-based approach to regulation, where the level of risk decides the level of regulation enforced on the technology. In the draft, AI is defined to include systems that can "generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with . . . for a given set of human-defined objectives." The law would prohibit the use or sale, as a general matter, of AI that:

- Manipulates the subconscious,
- Exploits vulnerabilities based on age or disability,
- Is used by government agencies to evaluate character and fitness using a social score, and/or
- Provides real-time biometric identification (with a few exceptions).

Although not meeting the factors for prohibition enumerated above, an AI system can be deemed "high risk" and therefore subject to strict regulations. These high-risk systems must have human oversight while in use.

Important to the purpose of this chapter is the draft law's section on governance. Titles VI, VII, and VIII specify the governance scheme, which requires the creation of a top-level board to assist the EC in providing guidance on AI and emerging issues: The European Artificial Intelligence Board. Supervisory authorities in individual countries are also expected, under the draft, to play a role as the board will be composed of the heads of each along with the European Data Protection Supervisor. The Member States' national authorities will be involved with the coordination of surveillance of AI markets, and enforcement of the regulations. This kind of network arrangement could be considered an example of lead-organization governance, where one organization coordinates major activities and decisions, making power asymmetrical and centralized.

In general, the AIA is an update of the EU's previous Coordinated Plan for AI that was first proposed in 2018. In particular, the goal of the AIA is to have member states collaborate to deal with the risks and harms possible from automated systems more effectively. In furtherance of this collaborative effort, the law requires the creation of a centralized database of high-risk AI systems that might impact fundamental rights. The creators of these systems are required, under the law, to register their creations and provide data to the

database. Further oversight of AI systems is facilitated through member state-based market surveillance authorities, responsible for investigating compliance with regulations for high-risk AI systems. Member states are also allowed to use their already-established sectorial authorities to assist in monitoring and enforcing the law.

Brazil: Artificial Intelligence Bill

In late September 2021, Brazil's House of Representatives approved Bill No. 21/20, which sets out a framework for the use and implementation of artificial intelligence (Heikkila, 2021). With requirements of impact assessments and transparency for public use of AI in the public use of algorithmic systems, the bill orders AI agents—those who develop and/or operate the technology—to ensure compliance with the law, as well as working to ensure responsible innovation. The proposed law offers guidelines for AI development and use, as well as promoting public sector use of AI to respond to societal challenges (Brazil Bill 21/20, 2021). Specifically, the law makes the beneficial purpose of the creation of the system and human-centered design some of the guiding principles for AI developers.

The Brazilian proposal differs from that of the EU in that there is no specific organizational governance scheme offered in the former. The bill does, however, provide guidelines for future regulation. Public entities are named as responsible for the creation of interventions specific to each sector. The adoption of norms that could impact the development and use of AI would only be made after wide-reaching public consultation. Public authorities will also be required to produce regulatory impact analyses. The bill also admonishes government entities at the union, state, federal, and municipal levels to promote trust in AI and offer incentives for AI research and development. At the same time, public entities at all levels of government are required to encourage the creation of regulatory instruments, regulatory sandboxes, transparency, and collaborative governance schemes in connection with business, civil society, and researchers. Public authorities are also asked to promote international cooperation and to attempt to harmonize legislation with the use of treaties and trade agreements. These requirements are designed to be executed by sectoral agencies in accordance with federal regulations.

United States: *AJOPT* and Algorithmic Accountability Act

An update of a bill proposed first in 2019, the U.S. Algorithmic Accountability Act (AAA) (2022), if passed, would force organizations to submit algorithmic impact assessments. These organizations would not only be those that create and develop AI systems, but also those organizations that are partnered with AI systems creators, assuming those organizations meet a regulatory threshold. Organizations are required to maintain the assessments for five years. These impact assessments would then be reported to the U.S. Federal Trade Commission, which is required under the bill to promulgate rules regarding organizational transparency and reporting. The bill also provides more funding for the FTC; it also mandates the creation of a Bureau of Technology within the Commission, and the expansion of the Bureau of Enforcement. Although providing for interagency agreements for cooperation, the proposed law places sole responsibility on the FTC as a governance

body. In contrast, *AJOPT*, as mentioned earlier, would initiate a cooperative effort among several federal agencies, while still allowing agencies on the state levels to work toward similar goals.

Complexity in action?

The AI-related proposals offer a glimpse into how policymakers around the world are attempting to create organizational governance schemes for AI at a national level. A major question remains as to whether these proposed schemes can match the complexity of the problematic use of AI systems in various sectors. According to Braithwaite et al. (2018), complexity, as a scheme, requires four elements:

1. A triggering mechanism;
2. Feedback loops;
3. Time; and
4. A systems-informed, complex approach that considers existing networks and socio-technical systems.

Complexity focuses on the relationships between components; the connections between agents and their artifacts are most important (Braithwaite et al., 2018). Agents adjust their behavior according to learning from each other and the environment. Behavior is dynamic, and local interactions between agents creates context and determines both present and future behavior. Local interactions create larger global, systemic patterns, as well as new factors. Patterns are then influenced by feedback loops (Braithwaite et al., 2018). Complexity, then, may be a useful way of considering the potential of organizational governance schemes to adequately make interventions in designated areas, like AI.

All three of the proposals above, and the *AJOPT* Act, meet the first requirement of a triggering mechanism. Legislation is considered an intervention into current schemes, which allows for strategic changes to happen. Likewise, all the previously discussed proposed legislation include feedback loops: the AIA with the cooperative board and transparency requirements, the Brazilian law with public comment, etc. Where these bills differ is in the fourth factor; only the AIA is explicit in its systems-informed approach that leverages already existing networks and socio-technical systems. While the AIA creates a lead organization network for administration and coordination, it is the participants in the network who are tasked with enforcement and who provide feedback for use by the larger Board, which may further decide on changes and implementations. This is an example of the local interactions influencing the broader global regulatory frames.

Although the Brazilian legislation provides only a framework for future policymaking, the language of the bill offers some insight into the use of multiple authorities for AI governance. The U.S. AAA, in contrast, places sole responsibility for rules and enforcement on one federal agency. Although the goal of algorithmic transparency is laudable, the focus on one federal agency may prove insufficient to meet these goals. The *AJOPT* Act, in contrast, appears to recognize the necessity of multiple entity collaboration to meet the goals protecting individuals and communities from algorithmic discrimination and unfairness. In sum, although all the proposed legislative actions offer measures of protection from AI

systems that can produce unfairness, few reflect the complexity of AI in the design of enforcement and organizational oversight. This reflects a simplistic approach to intervention into the overall governance system, which may not lead to the desired results because it fails to recognize that relationships between agents offer the opportunity for agile enforcement, knowledge sharing, and, ultimately, course correction to meet the goals of statecraft.

Designing Complex Organizational Governance

Organizational governance schemes that look to the power of a network of actors while recognizing the need for coordination and taking a real-world perspective on the nature of a problem offer a more adequate way of governing complex systems. Networks allow for strategic deployment of enforcement, knowledge sharing, and task management. Network effectiveness, or "the attainment of positive network-level outcomes that could not normally be achieved by individual organizational participants acting independently" (Provan & Kenis, 2008, p. 230), is linked to competencies necessary to complete the tasks applicable to meeting the goals of the network (Provan & Kenis, 2008). Therefore, the involvement of organizations, or agencies, with different subject matter expertise should positively influence the network effectiveness. A regulatory scheme that leverages the power of networks and complexity, then, might include the following:

1. *A network of several agencies with different subject matter expertise*. The involvement of several government level agencies working in concert to match the complexity of an issue like AI and the use and deployment of algorithmic decision systems would allow for better enforcement of law. Although AI is the technology, or set of technologies, in need of regulation, AI and its uses and impacts can appear differently in different contexts. In housing, AI might look, in one instance, like the use of facial recognition technology for entry into low-income housing projects. In healthcare, for example, the use of AI might look like the use of algorithms to provide guidance about who should be placed on transplant lists, and where on the list an individual should be placed. While these are both examples of uses of AI, the different contexts influence the different kinds of impacts. It is important, then, for representation from agencies that have differing contextual expertise for more adequate and holistic governance. Each agency will be tasked with promulgating rules and programs designed to meet the goals of the legislation.
2. *An independent administrative or coordinator agency*. Along with the network of agencies with different expertise, it will be important to have a separate organization or agency that will administer and coordinate activities and assignments related to regulation and enforcement. With this element, the network power is dispersed among the separate organizations, with the separate administrative organization brokering the network. This agency, then, will ensure that the other networked agencies are coordinating their efforts in service of the larger goals; in this case, the regulation of uses and mitigation of harms from AI technologies.

3. *Feedback loops of consultation and revision for regulation*. Because organizational networks are akin to "a collection of programs and services that span a broad range of cooperating but legally autonomous organizations," (Provan & Milward, 2001, p. 417), it will be important to continually assess the effectiveness of the network and to make changes when necessary. This requires the creation of opportunities from the communities and organizations affected by the network. This might take the form of public comment, town halls, and evaluations of the network through comments from agency representatives. These assessments should then be used for reevaluation and revision of the network and network function.

The three requirements enumerated here may appear simplistic considering the complexity of AI and the creation of regulation that adequately addresses the potential harms from the various algorithmic technologies. These requirements are designed to intervene into an area of concern by strategically using the power of a network to regulate and mitigate harms. The network, with coordination, can impact the nodes of concerns with the technology in various contexts because of differing agency expertise. More than this, the network removes the burden on one agency from having to attempt to regulate a multifaceted and evolving technology. In addition, the feedback loops allow for regulation to be iterative and more effective if agencies use the information gained for revision.

Conclusion

This chapter has offered a brief examination of the possibilities of making the organizational aspect of governance, as offered through legislation, more complex. Complexity offers a way of approaching the "real-world, multidimensional appreciation of the system and its density and dynamics" (Braithwaite et al., 2018), as reflected in AI and algorithmic systems. Adding elements of complexity, specifically appreciating the networked nature of technology, sectors, and outcomes, can assist in furthering the goals of interventions, such as legislation. Creating policy that recognizes the utility of networked governance, in which multiple entities coordinate to govern a subject area, offers solutions for policy that more adequately reflects the reality of the networked relationships in AI systems. As more governments take up the cause of attempting to regulate AI, it is worthwhile for policymakers to recognize the governance potential of networked organizations and how adding this to legislative interventions can assist in matching the complexity of AI governance. Of course, the use of a more complex system of governance does not guarantee change. It does, however, offer a model of governance that focuses on the reality of an influential and increasingly ubiquitous range of systems.

References

Alter, C., & Hage, J. (1993). *Organizations working together*. Sage Publications.

Bayrakal, S. (2006). The U.S. Pollution Prevention Act: A policy implementation analysis. *The Social Science Journal 43*(1), 127–145. https://doi.org/10.1016/j.soscij.2005.12.012.

Bevir, M. (2012). *Governance: A very short introduction*. Oxford University Press.

Estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá outras providências., no. PL n. 21/2020 (2021). https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=1853928.

Braithwaite, J., Churruca, K., Long, J. C., Ellis, L. A., & Herkes, J. (2018). When complexity science meets implementation science: A theoretical and empirical analysis of systems change. *BMC Medicine 16*(1), 63. https://doi.org/10.1186/s12916-018-1057-z.

Brass, D. J., Galaskiewicz, J., Greve, H. R., & Tsai, W. (2004). Taking stock of networks and organizations: A multilevel perspective. *Academy of Management Journal 47*(6), 795–817. https://doi.org/10.5465/20159624.

Cole, D. H., & Grossman, P. Z. (2018). When is command-and-control efficient? Institutions, technology, and the comparative efficiency of alternative regulatory regimes for environmental protection. In G. E. Helfand & P. Berck (Eds.), *The Theory and Practice of Command and Control in Environmental Policy* (1st ed., pp. 115–166). Routledge. https://doi.org/10.4324/9781315197296-7.

Cox, M. (2016). The pathology of command and control: A formal synthesis. *Ecology and Society 21*(3). https://www.jstor.org/stable/26269979.

Dhar, P. (2021). AI legislation must address bias in algorithmic decision-making systems. *VentureBeat*, July 5. https://venturebeat.com/2021/07/05/ai-legislation-needs-to-broaden-its-focus-from-rd-to-address-bias-in-algorithmic-decision-making-systems/.

Gereffi, G., & Mayer, F. (2006). Globalization and the demand for governance. In G. Gereffi (Ed.), *The new offshoring of jobs and global development*, 39–65.

Goodfellow, T., & Lindemann, S. (2013). The clash of institutions: Traditional authority, conflict and the failure of "hybridity" in Buganda. *Commonwealth & Comparative Politics 51*(1), 3–26. https://doi.org/10.1080/14662043.2013.752175.

Heikkila, M. (2021). Brazil's AI law—US takes a risk-based approach—Social scoring. *Politico*, November 17. https://www.politico.eu/newsletter/ai-decoded/brazils-ai-law-us-takes-a-risk-based-approach-social-scoring/.

Hudson, B., Hunter, D., & Peckham, S. (2019). Policy failure and the policy-implementation gap: Can policy support programs help? *Policy Design and Practice 2*(1), 1–14. https://doi.org/10.1080/25741292.2018.1540378.

Li, H., Sarathy, R., & Zhang, J. (2010). Understanding compliance with internet use policy: An integrative model based on command-and-control and self-regulatory approaches. *Proceedings of the 31th International Conference on Information Systems*, Paper 181.

McNealy, J. E., Jen, D., & Nguyen, S. (2022). Prototyping policy: Visualizing impact for better regulation. *Convergence*, 28(1), 109–126. https://doi.org/10.1177/13548565211069875

Pierre, J., & Peters, B. G. (2019). *Governance, politics and the state*. Red Globe Press.

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, no. COM (2021) 206 final, European Parliament (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

Provan, K. G., & Kenis, P. (2008). Modes of network governance: Structure, management, and effectiveness. *Journal of Public Administration Research and Theory 18*(2), 229–252. https://doi.org/10.1093/jopart/mum015.

Provan, K. G., & Milward, H. B. (2001). Do networks really work? A framework for evaluating public-sector organizational networks. *Public Administration Review 61*(4), 414–423.

Pylant, A.C. (2020). Initiating a Collaborative Cybersecurity Governance Framework at the State Level. [Doctoral Dissertation, West Chester University] West Chester University Digital Commons. https://digitalcommons.wcupa.edu/cgi/viewcontent.cgi?article=1083&context=all_doctoral.

Robichau, R. W. (2011). The mosaic of governance: Creating a picture with definitions, theories, and debates. *Policy Studies Journal 39*(s1), 113–131. https://doi.org/10.1111/j.1541-0072.2010.00389_8.x.

Senator Markey, Rep. Matsui Introduce Legislation to Combat Harmful Algorithms and Create New Online Transparency Regime. (2021, May 27). Ed Markey, U.S. Senator for Massachusetts. https://www.markey.senate.gov/news/press-releases/senator-markey-rep-matsui-introduce-legislation-to-combat-harmful-algorithms-and-create-new-online-transparency-regime.

Sinclair, D. (1997). Self-regulation versus command and control? Beyond false dichotomies. *Law & Policy 19*(4), 529–559. https://doi.org/10.1111/1467-9930.00037.

Sørensen, E., & Torfing, J. (2009). Making governance networks effective and democratic through metagovernance. *Public Administration 87*(2), 234–258. https://doi.org/10.1111/j.1467-9299.2009.01753.x.

Sørensen, E., & Torfing, J. (2016). *Theories of democratic network governance*. Springer.

Stivers, C. (2008). *Governance in dark times: Practical philosophy for public service*. Georgetown University Press.

Stoker, G. (2006). Public value management: A new narrative for networked governance? *The American Review of Public Administration 36*(1), 41–57. https://doi.org/10.1177/0275074005282583.

Zakrzewski, C. (2021). Democrats introduce legislation prohibiting algorithmic discrimination. *Washington Post*, May 27. https://www.washingtonpost.com/politics/2021/05/27/technology-202-democrats-introduce-legislation-prohibiting-algorithmic-discrimination/.

CHAPTER 30

THE ROLE OF WORKERS IN AI ETHICS AND GOVERNANCE

NATALIYA NEDZHVETSKAYA AND J. S. TAN

Introduction

AI is a social technology. It is designed by teams of developers who are employed by corporations, governments, academic institutions, or non-profit organizations. Using massive amounts of existing data, the technology learns from what it is fed (Valiant, 1984). Once implemented, it reflects—and automates—the biases that come both from its designers and the data it was trained on (Denton et al., 2020). Without careful consideration of how this technology is designed and integrated into the systems we use, it can reproduce the biases and inequities that occur in everyday life (Friedman & Nissenbaum, 1996; Buolamwini & Gebru, 2018; Noble, 2018; Benjamin, 2019). As AI systems get deployed at an ever faster rate, affecting millions of people around the world, the scale at which such harms are propagated can be enormous (Crawford, 2021). Communities around the world—consumers, researchers, activists, and workers—have taken steps to understand and counteract the repercussions of these technologies on their lives. The field of AI governance has grown in tandem with AI technology with the goal of designing systems and protocols to minimize the potential harm of AI technologies (Dafoe, 2018).

While the role of states, corporations, and international organizations in AI governance has been extensively theorized, the role of workers in AI governance has received comparatively little attention. This chapter looks at the role that workers play in defining and mitigating harm that comes from AI technologies. We focus on societal harms as a category distinct from system reliability failures (Lynch & Veland, 2018). Harms are the causally assessed "impacts" of technologies (Moss et al. 2021). They are not a result of technical negligence or a lack of oversight, but rather of normative uncertainty around questions of safety and fairness in complex social systems (Altman et al., 2018; Raji et al., 2020; Dobbe et al., 2020). Numerous AI ethics frameworks conceptualize workers as stakeholders with a professional and personal interest in upholding ethics and preventing harms (Jobin et

al., 2019). Within workplaces, however, the ability of workers to act as stakeholders is dependent on a number of factors. Examining the past decade of AI-related worker activism allows us to understand how different types of workers are positioned in workplaces that produce or propagate AI systems, how their position informs their claims, and the place of collective action in staking these claims.

Over the past decade, organizations that produce AI systems have sought to address these harms by establishing internal divisions for AI governance (cf. Google, 2021; Microsoft, 2021). These divisions operate as an internal reporting channel through which workers can report harms (Rakova et al., 2021). They also contribute to a growing professional field of AI ethics that generates frameworks for diagnosing harms and treatments for addressing them. In some cases, these internal structures function as expected, and reports of harm are acknowledged and handled within the workplace. However, in other notable cases, workers and management disagree on how harms are identified or treated and workers turn to collective action to stake their claim on questions of AI governance (Belfield, 2020; Tarnoff, 2020).

In this chapter, we construct a model of harm reporting practices to explain how and why workers turn to collective action as a response. We draw on archival data of 25 collective actions publicly reported in the U.S. AI industry from 2010 to 2020 to inform our model (Tan & Nedzhvetskaya, 2022). We divide the harm reporting process into three steps (identification, governance decision, and response) with three stakeholders in the process (management, AI workers, and AI ethicists). Notably, our definition of AI workers includes both "designers," workers who design, construct, or apply an AI technology or its surrounding infrastructure, and "trainers," workers who feed data into an existing AI technology, supplement AI systems with human labor, or whose work is controlled by AI systems. Our archival data also shows that workers draw upon three types of claims to call for jurisdiction over questions of AI governance: subjection, control over the product of one's labor, and proximate knowledge of systems. Workers turn to collective action to challenge higher-level decisions in the identification, governance decision, and response processes. Through collective action, workers lay claim to their own perspective on AI governance in the workplace. We examine these claims in later sections of the chapter and evaluate how they may differ across categories of AI workers.

Who Is an AI Worker? Who Is an AI Ethicist?

To explore the role of workers in AI governance, it is necessary to begin with a definition of an AI worker. An AI worker must be employed or contracted by an institution that produces or uses AI systems. This could include public or private companies, academic institutions, and non-profit organizations. AI workers fall into two categories: "designers" and "trainers".[1] The first category, "designers," are workers who design, construct, or apply an AI technology or its surrounding infrastructure. This includes AI researchers, as well as those involved in developing and deploying AI systems, such as computer scientists and programmers. The second category, "trainers," are workers who feed data into an existing

AI technology, supplement AI systems with human labor, or whose work is controlled by AI systems. This includes workers who are reimbursed for the data they provide to train AI technologies—for example, rideshare drivers or app-based shoppers—and "artificial artificial intelligence" workers who supplement AI systems that would otherwise fail to perform "intelligently" (Shestakofsky & Kelkar, 2020).[2]

The labor of "trainers" is as important to the design and function of algorithms as the code written by "designers," and this has led to critiques of this binary (Amrute, 2019). We agree with this critique but choose to use the categories of "designer" and "trainer" because their entanglement with power dynamics explains some of the patterns we see in how these two categories of workers are treated across workplaces. This, in turn, shows why certain workers are more likely to make certain claims to AI governance. For example, AI trainers are more likely to be subject to harm from the algorithm itself and to use this claim when organizing a collective action. These two categories are entangled with power dynamics within institutions and within society more generally. Perceived value on the labor market can translate to differential treatment within the workplace (Rosenblat, 2018; Belfield, 2020). In many workplaces, divisions between these two categories of workers reflect existing inequalities in society along the lines of socio-economic class, race, gender, and nationality (Benjamin, 2019; Gray & Suri, 2019; Jack & Avle, 2021).

Another group of stakeholders that we must consider are AI ethicists: those involved in understanding and critiquing AI systems with the goal of reducing AI harms. While an AI ethicist may share some of the skills that AI workers have, such as implementing or deploying AI systems, they focus their time not on implementation, but on researching these systems for their potential to cause harm. Their expertise is not solely the result of professionalization (e.g., specialized training or professional designations). Rather, expertise is a complicated arrangement of abstract and experiential knowledge, social ties, and institutional legitimacy (Eyal, 2013). AI ethicists spend time in different spaces (e.g., ethics conferences), establish different social ties (e.g., with fellow AI ethicists across companies, research centers, and universities), accumulate different experiences, and bear different institutional pressures than AI workers. Thus, AI ethicists form their own "epistemic community" with a shared set of experiences distinct from AI workers (Haas, 1992). Nevertheless, internal AI ethicists are employees in the same respect that AI workers are employees and face similar pressures to prove their value to their employer and comply with the norms and goals of their institution.

Institutions have many reasons for hiring experts such as AI ethicists. Expertise is a claim to jurisdiction over some field of abstract knowledge associated with a set of concrete tasks (Eyal, 2013). Institutions seek to ensure their own survival and reflect this in their actions (Berger & Luckmann, 1966; Meyer & Rowan, 1977). Experts can bring legitimacy to an institution, increasing their chances of survival by strengthening support for their actions both internally and externally. Notably, experts are valued not for their ability to apply objective facts in uncontested situations but rather for their ability to use complex inference to solve problems in contested domains (Abbott, 1988). When institutions hire experts for their expertise and retain oversight of their work, they maintain jurisdiction for the institution itself to the extent that this jurisdiction does not encroach on external professional or legal codes. For example, an institution may hire a lawyer for their legal expertise. The lawyer will work on behalf of the institution but within the realm of legal code. By hiring internal

ethicists, institutions maintain control over how the abstract knowledge of expertise can be applied to critique the institution's operations.

This is especially the case in the field of AI ethics. Because of the nature of the field, AI ethicists' expertise extends not only to empirical knowledge but also to moral questions, which can make them a crucial resource for institutions seeking legitimacy for their AI products and services (Bostrom & Yudkowsky, 2011; Greene et al., 2019). As with other forms of business ethics, AI ethics provides a moral background for the field of AI, a set of causally informed moral concepts that allow us to establish norms that inform our behaviors (Abend, 2014). Institutions hire AI ethicists for their jurisdiction over what is contested, uncertain, and subjective when it comes to AI harms. In the following section, we demonstrate how ethicists are integrated in AI governance in the workplace and the ways in which they must simultaneously alternate between their roles as experts and workers. We begin by introducing the data that informed our model before explaining the harm reporting process in AI workplaces.

Harm Reporting in AI Workplaces

Our model of harm reporting practices is limited to workplaces that produce AI systems, or what we will call an "AI workplace" from here on. To construct this model, we draw upon archival data on collective actions in the technology industry from 2010 to 2020. Of the over 300 actions catalogued, 25 directly mention or involve AI. Table 30.1 provides a chronological list of these collective actions. From here on, we refer to these as *AI collective actions*. In our harm reporting model, we divide the institution's process for evaluating and managing harm into three parts—identification, governance decision, and response—and identify the roles played by management, AI workers, and AI ethicists (Table 30.2). Notably, this framework applies to both "designers" and "trainers."

To construct our archive on collective actions in the technology industry, we gathered data using NexisUni news archives. We searched for articles where collective action terms[3] occurred within 25 words of employment terms[4] for the computing and information technology industry. To maintain the archive, we draw from a variety of sources including, but not limited to, a network of active organizers in the tech industry, reporters on the tech-labor beat, and Google Alerts. To qualify for our archive, events must be "collective"[5] and present "evidence of action"[6] by currently or recently employed "tech workers."[7] For this analysis, we only include actions that directly mention or involve AI, protesting the use of an AI technology, the application of AI to govern or control their own labor, or the stakes of AI governance in the workplace.

Our archive of events is limited only to actions that have been reported by a news publication and is heavily skewed through actions reported in the English language press because we conducted the search only in English. In addition, news publications can be biased in their reporting of social movement actions based on their geographic locale and political orientation (Davenport, 2009). We believe this to be the case for collective actions in the technology industry as well. Seventy-two percent of actions covered in our archive through 2020 occurred in the United States, Canada, or online and 84 percent were in the United

Table 30.1 Public Collective Actions in AI, 2010 to 2010

Date	Institution	Collective Action	#	Claims	Worker Type	Description
Feb. 2013	Amazon	New Platform	NA	Subjection; Proximate Knowledge	Trainers; Designers	Mechanical Turk workers have adopted a platform named Turkopticon, which allows them to review tasks and requesters on the platform.
Sept. 2014	Amazon	Open Letter; Standard Setting	262	Proximate Knowledge; Product of One's Labor; Subjection	Trainers; Designers	Mechanical Turk workers and academics, organized through Dynamo, have collaborated and signed a set of ethical guidelines for scholars requesting labor through the platform.
Dec. 2014	Amazon	Open Letter	550	Subjection	Trainers	Mechanical Turk workers participate in an email writing campaign to Amazon founder Jeff Bezos to protest low pay and worker representation.
Aug. 2017	Amazon	New Platform	550	Proximate Knowledge; Product of One's Labor; Subjection	Trainers; Designers	Mechanical Turk workers and organizers collaborated with researchers at Stanford's Crowd Research Collective to create Daemo, a crowd-sourced platform that provides a higher-paying alternative to Amazon's Mechanical Turk.
April 2018	Academic Institutions	Open Letter	50	Product of One's Labor; Proximate Knowledge	Designers; Ethicists	A group of academics working on artificial technologies call for a boycott of the Korea Advanced Institute of Science and Technology, a South Korean university believed to be developing AI weapons through a collaboration with a defense contractor.
May 2018	Google	Open Letter	5000	Product of One's Labor	Mostly Designers	Google employees have led a campaign demanding that their company terminate its contract with the Pentagon for Project Maven, a program that uses machine learning to improve targeting for drone strikes.
June 2018	Google	Internal Protest	9	Product of One's Labor	Designers	A group of influential software engineers in Google's cloud division, referred to as the "Group of Nine," surprised their superiors by refusing to work on a cutting-edge security feature. Known as "air gap," the technology would have helped Google win sensitive military contracts.

July 2018	Microsoft	Open Letter	500	Product of One's Labor	Mostly Designers	Two Microsoft employees presented their CEO with an online petition including more than 300,000 signatures (including 500 employees) calling for the firm to cancel their contract with the U.S. Immigration and Customs Enforcement Agency.
Oct. 2018	Microsoft	Open Letter	NA	Product of One's Labor	Designers	An open letter signed by "employees of Microsoft" asks the cloud giant to abstain from bidding on the military's massive JEDI cloud computing contract for ethical reasons involving the application of artificial intelligence.
Feb. 2019	Microsoft	Open Letter	290	Product of One's Labor	Designers	Microsoft employees circulated a letter among the companies' over 130,000-person staff demanding that executives cancel a $479 million contract with the U.S. Army, IVAS, that will provide weapons technology to the U.S. Military to "increase lethality" to army soldiers.
March 2019	Amazon	Open Letter	45	Product of One's Labor	Designers; Ethicists	Researchers are calling on Amazon to stop selling facial recognition software to law enforcement.
April 2019	Google	Open Letter	2556	Product of One's Labor	Mostly Designers	Google employees, along with academic, civil society, and industry supporters, have called for the removal of a rightwing thinktank leader from the company's new artificial intelligence council, citing her anti-LGBT and anti-immigrant record.
May 2019	Facebook	Internal Protest	12	Subjection	Trainers	Facebook moderators in the United States have been spearheading a quiet campaign inside the social media giant to air their grievances about unsatisfactory working conditions and their status as second-class citizens.
Nov. 2019	Instacart	Open Letter; Protest	212	Subjection; Proximate Knowledge	Trainers	Instacart workers circulated an open letter on Medium prior to a walkout. Employees are protesting the company's history of systematically devaluing labor, using algorithms to reduce pay to employees and changing tipping structures to return profit to the company.

(*continued*)

Table 30.1 Continued

Date	Institution	Collective Action	#	Claims	Worker Type	Description
Nov. 2019	Google	Open Letter	1137	Product of One's Labor	Designers; Ethicists	Google employees have signed a letter calling for their employer to take action on the issues of climate change and immigration. The letter called on Google to commit to zero carbon emissions by 2030, end all contracts with fossil fuel companies and climate change denying or delaying "think tanks, lobbyists, and politicians," and end all collaboration with organizations or individuals "enabling the incarceration, surveillance, displacement, or oppression of refugees or frontline communities."
2019	Meituan, Ele.me, Fengniao	Strikes	NA	Subjection	Trainers	In 2019, food delivery drivers have participated in over 45 strikes across China over issues including demanding higher pay and protesting change in distance calculation methods
Feb. 2020	Academic Institutions	Open Letter	141	Product of One's Labor	Designers; Ethicists	AI researchers and engineers from around the world signed an online open letter calling for the tech industry to refrain from using AI technologies to "exacerbate the climate crisis."
June 2020	Google	Open Letter	1650	Product of One's Labor	Designers	Google employees have signed an open letter to CEO Sundar Pichai demanding the company stop selling its technology to police forces across the U.S.
June 2020	Academic Institutions	Open Letter	600	Proximate Knowledge	Designers; Ethicists	Hundreds of expert researchers and practitioners across a variety of technical, scientific, and humanistic fields (including statistics, machine learning, artificial intelligence, law, sociology, history, communication studies, and anthropology) sign a letter calling for a forthcoming publication entitled "A Deep Neural Network Model to Predict Criminality Using Image Processing," to be rescinded from publication.

June 2020	Microsoft	Open Letter	250	Subjection	Designers	Microsoft employees wrote and collectively signed an email to Microsoft's leaders calling for the company to cancel its contract with the Seattle Police Department. In the following month, Microsoft banned the use of their facial recognition software.
Oct. 2020	Facebook	Open Letter	290	Subjection	Designers; Trainers	Facebook employees and contractors signed a petition demanding a 50 percent wage increase as hazard pay for site moderators who have been forced to return to the office by October 12, 2020. A Facebook spokesperson said the reason for the return was that "some of the most sensitive content can't be reviewed from home."
Oct. 2020	Uber	Legal Action	1000	Subjection	Trainers	British Uber drivers are launching a lawsuit in the Netherlands to protest the company's use of algorithms to fire workers.
Nov. 2020	Facebook	Open Letter	200	Subjection	Trainers; Designers	Facebook workers signed a letter demanding better treatment for content moderators after being required to return to the office in the midst of the pandemic.
Dec. 2020	Google; Academic Institutions	Open Letter	2695	Product of One's Labor; Proximate Knowledge	Designers; Ethicists	After Dr. Timnit Gebru was fired from Google, Googlers wrote an open letter demanding accountability from Google.
Dec. 2020	Google	Open Letter	NA	Product of One's Labor; Proximate Knowledge	Ethicists	Two weeks after Dr Timnit Gebru was fired from her role in Google, Google's Ethical AI employees sent CEO Sundar Pichai a list of demands, including organizational changes and a request to reinstate dismissed researcher Timnit Gebru at a higher level.

Table 30.2 Harm Reporting Process Model

1. Identification	2. Governance Decision	3. Response
	Consensus on governance framework	Prescribed treatment
	No consensus on governance framework	Collective action
		Individual action (e.g. whistleblowing)
		No action

States, Canada, Europe, or online. Fifty-eight percent of actions took place in one of seven American corporations: Amazon, Apple, Facebook, Google, Lyft, Microsoft, and Uber. We have attempted to counteract these biases by opening the archive to crowdsourced contributions, which are more likely to come from outside of the United States and to include smaller companies and institutions. To date, approximately five percent of events have been contributed to the archive through crowdsourcing. In the following sections, we draw upon cases from Table 30.1 to illustrate how harm reporting practices have proceeded in AI workplaces.

Identification

During *identification*, harms are identified by AI workers as part of their routine labor or by AI ethicists as part of an audit or research study.[8,9] The identification process can take place immediately upon encountering the harm or gradually over the process of becoming acquainted with the systems at work. Identification may preempt potential harms, as in the case of Microsoft employees that called on the company to abstain from bidding on the now-cancelled Joint Enterprise Defense Infrastructure (JEDI) contract from the U.S. Department of Defense (Employees of Microsoft, 2018). The JEDI contract is designed to accelerate the military's adoption of AI technology (Williams, 2020) and increase the military's lethality (Employees of Microsoft, 2018).

In other instances, harms have been identified years after their initial repercussions were felt. In 2020, British Uber drivers filed a lawsuit in Dutch courts alleging that the company's firing algorithm violated GDPR Article 22. More than one thousand individual cases of firings were presented, dating two years back (Russon, 2020). While individual workers identified the harms much earlier, when they presented these harms to the company, the company failed to recognize them as such. By gathering cases together in a class action lawsuit, that is, by taking a collective action, workers were able to contest this decision. Among the two categories of AI workers described, trainers are more likely to find themselves the subject of harm. The nature of the work is itself often governed by algorithms and workers are often the first to identify harms as a result (Rosenblat & Stark, 2016).

Governance decision

Once a harm is identified, the question of how to manage that harm, and who gets to decide, comes next. This leads us to the second step of the model, the *governance decision*. The *governance decision* is a way for the institution to decide how to manage an identified harm. Typically, decisions made in this section are controlled by management. Here, management must evaluate: is there an existing governance framework that determines how this harm should be dealt with? During the governance decision, management may consult with AI ethicists to see if there is an existing framework to diagnose the identified harm.

Conflicts can arise during the governance decision when recognizing a harm threatens the survival of an institution or reduces its legitimacy. Corporations, for instance, must demonstrate their present or future potential for profit to receive investment from their shareholders (Useem, 1993; Fligstein, 2001). When reported harms threaten the success of profitable products, managers within a corporation must decide between their obligation to generate profit for shareholders or their obligation to adhere to the ethical standards of the field. The two may coincide if management believes that a harm poses enough of a reputational risk to the firm that it threatens future profitability (Rakova et al., 2021).

Among the 25 collective actions documented, 10 protested specific corporate contracts. Conflicts between corporate pressures to generate profit and workers' personal and professional ethical standards are a common source of collective action in the AI workplace. In 2020, Dr. Timnit Gebru, a co-lead on Google's Ethical AI Team, was removed from the company after submitting an academic paper to management for review. The paper included critiques of a category of language processing models that are core to the company's highly profitable search engine business (Hao, 2020). The conditions of Dr. Gebru's removal from the company led individuals close to the case to allege that the action was a case of censorship and incited a number of collective actions from Google workers and the broader AI research community (Google Walkout for Real Change, 2020; Canon, 2020; Allyn, 2020).

Notably, concerns over lack of legitimacy triggered by worker collective actions can lead management to formalize ethics review processes, laying out objectives and bolstering AI ethicists on staff. In June 2018, following several months of internal activism against the company's involvement with Project Maven, an AI program contracted by the U.S. Department of Defense, Google CEO Sundar Pichai released a blog post announcing the company's new AI principles (Shane et al., 2018; Pichai, 2018). In the post, Pichai stated that the company would not "design or deploy AI . . . that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints." Formalizing review processes and articulating ethical principles can improve accountability, but does not address the issue of oversight and objectivity in internal governance without accompanying organizational change (Posada, 2020; Amrute, 2021; Rakova et al., 2021).

Response

The *response* can take a few forms. When a framework exists to diagnose the harm and there is high consensus among all stakeholders, the harm can be treated through the prescribed

method: management consults with AI ethicists, and AI ethicists are able to diagnose and treat the harm. Treatment of the harm can also involve AI workers. "Trainers" may have to re-label data or collect data in a different way to mitigate the identified harm. "Designers" may have to alter the nature of the product itself. Treatment itself can generate conflict, however, if stakeholders disagree about its application or feel it is insufficient to address the scale of the harm. Lack of consensus can occur when management disagrees with AI ethicists' diagnosis of the harm or dismisses or reclassifies the harm. In cases where an AI ethicist is unable to exercise their expertise, they must participate as a worker to stake their claim.

If there is low consensus among stakeholders about the existing framework or its prescribed treatment, AI workers may choose to engage in individual action (e.g., whistleblowing) or collective action as part of the response. Collective action is defined by social movement theorists as "emergent and minimally coordinated action by two or more people that is motivated by a desire to change some aspect of social life or to resist changes proposed by others" (McAdam, 2015). While this definition is broad enough to encompass all stakeholders involved in our model—management, AI workers, and AI ethicists—collective action is seldom taken by management because managers have enough power within the institution to act individually. For AI workers, what is crucial is that the power to incite or resist change is partly the result of acting as a collective claiming to represent a viewpoint greater than the individual.

AI workers are more likely to engage in collective action when other means of influence are not available to them or have proven insufficient. Through collective action they can contest an identification process or a governance decision and gather support, either internally within the company or externally with a broader public, for a different outcome. Critically, such actions do not come without consequences, and there have been numerous accounts of retaliation towards AI workers that have spoken out against their employers.[10] Fundamentally, collective actions in the workplace are political in that they challenge existing power structures in the workplace. Importantly, AI ethicists, as well as AI workers, may engage in collective action. We will next examine some of the ways in which workers make these claims.

Worker Claims

AI workers turn to collective action to challenge higher-level decisions in the identification, governance decision, and response processes in the AI workplace. Through collective action, AI workers stake their own claims on questions of governance. In the AI collective actions we have identified, there are three primary ways in which AI workers construct their claims to governance: subjection, control over the product of one's labor, and proximate knowledge of a system (Table 30.3).

First, AI workers themselves may claim to be subject to the harms they identify. In the case of delivery workers whose work is controlled by AI-powered algorithms, workers themselves experience the harms of exploitative systems that are designed to extract the greatest possible amount of labor out of employees, for example, incentivizing them towards overwork (Rosenblat & Stark, 2016; Lei, 2021). Delivery workers in China offer a

Table 30.3 Types of Worker Claims

Claim	Description	AI Worker Types	Example
Subjection	Workers themselves are subject to the harms they identify within AI systems	Trainers more vulnerable to subjection	British Uber drivers filed a lawsuit in the Netherlands to protest the company's use of algorithms to fire workers (Oct. 2020).
Product of One's Labor	Workers contest how the products of their own labor are being applied	Both types of workers may claim	An open letter signed by "employees of Microsoft" asks the company to abstain from bidding on the military's massive JEDI cloud computing contract for ethical reasons involving the application of AI to warfare (March 2019).
Proximate Knowledge	Workers claim a privileged insight into the workings and consequences of the AI systems they create by virtue of their proximity	Both types of workers may claim	Mechanical Turk workers and academics, organized through Dynamo, have collaborated and signed a set of ethical guidelines for scholars requesting labor through the platform (Sept. 2014).

good example of AI workers identifying and resisting AI-powered algorithmic management. Ele.me and Meituan, the two largest food delivery platforms in China, each have their own AI systems that, according to their creators, reduce the time for deliveries and give them an advantage over competitors.[11] However, from the perspective of the delivery drivers, these systems simply pass the burden of making more efficient deliveries onto drivers (*Renwu Magazine*, 2020). Drivers have demonstrated how the apps instruct them to take more dangerous routes to make a delivery on time, for example, advising them to ride against traffic on a one-way street. In response, delivery workers have participated in dozens of strikes, directly or indirectly protesting the harms they have faced because of the algorithm. According to the China Labor Bulletin, there were 45 recorded strikes in 2019 by food delivery drivers in China.[12]

Notably, AI workers may also identify with communities that are disproportionately impacted by the harms they identify, for example, vulnerable immigrant populations or racial targets of incarceration who are more likely to be subject to surveillance technologies (Selbst, 2017; Benjamin, 2019). Many of the leaders of collective actions in the technology industry have been women, people of color, or have identified as LGBTQ (Tarnoff, 2020). This is particularly salient for AI designers who themselves may not be the subject of harm but identify with particular communities or social movements. During the Black Lives Matter movement in 2020, many tech workers who identified as activists in the broader movement for racial justice raised the issue of police brutality and surveillance in their workplaces. Microsoft employees—who drew on their own experiences and participation in the movement—wrote and collectively signed an email to Microsoft's leaders calling for the company to cancel its contract with the Seattle Police Department (SPD). "Every one of us in the CC line are either first hand witnesses or direct victims to the inhumane responses

of SPD to peaceful protesting" the email states (Gershgorn, 2020). While the explicit demand to cancel the contract with the SPD was ignored, Microsoft banned the use of their facial recognition software the following month.

Second, AI workers may contest how the products of their own labor are being applied. They may identify sufficiently with the product of their labor that they feel responsible for, determining that it goes towards an end they view as beneficial or optimal (Caves, 2000; Ranganathan, 2018). In the aforementioned 2018 open letter by Microsoft employees asking management to drop a bid for the Department of Defense's JEDI contract, employees drew comparisons across ethics-driven activism across technology firms. "Like those who took action at Google, Salesforce, and Amazon, we ask all employees of tech companies to ask how your work will be used, where it will be applied, and act according to your principles," the letter stated (Employees of Microsoft, 2018). Such themes reverberate in other open letters, particularly those involving social and ethical issues, such as climate change (Amazon Employees for Climate Justice, 2019; Tech Won't Drill It, 2020; Microsoft Employees, 2019).[13]

Notably, both AI designers and AI trainers may claim control over the product of one's labor. In one documented incident, AI trainers raised concerns after they were asked to tag photos of home interiors that featured people who were naked or in the process of changing clothes (Posada, 2021). The trainers spoke out on an internal message board, raising concerns that the individuals featured in the photos might not be aware of the surveillance and that such work could be considered unethical. Benchmark machine learning datasets—and the trainers who have coded them—play a foundational role in determining how AI systems understand abstract social concepts. When the data itself is problematic, e.g., featuring violent, discriminatory, invasive, or objectifying images or language, the resulting algorithms can reflect these same problems. The biases and structural limitations of these datasets create epistemic commitments—"genealogies of data"—that proliferate as adoption of these benchmark datasets grows (Denton et al., 2020).

Third, AI workers may have a privileged insight into the workings and consequences of the products they create by virtue of the time they spend constructing them; that is, they have proximate and privileged knowledge of the systems (Perrow, 1984). In 1986, a group called Computer Professionals for Social Responsibility (CPSR), which included 14 AT&T employees and 30 external computer scientists from industry and academia, wrote a letter to U.S. President Ronald Reagan alleging that "Star Wars," a software system intended to defend the U.S. against ballistic missiles, was too error-prone to responsibly run (Boffey, 1986). Much of the group's argument drew upon their combined expertise in software development. They cited the inevitability of "bugs" in software code, the difficulty of fully testing software built at such a large scale, and the system's reliance on tightly coupled feedback loops.[14]

While AI technology was not implemented as part of the "Star Wars" system in 1986, the focus on system complexity and safety has been a consistent theme in studies of AI safety and AI accidents (Amodei et al., 2016; Maas, 2018). AI accidents are defined as "harmful behavior[s]" that result from a misspecification, a misunderstanding, or an oversight of AI processes (Amodei et al., 2016). AI accidents can be considered part of a larger genre of "system accidents," also called "normal accidents," which are the result of mistakes that rapidly escalate as a result of tight coupling and feedback loops in systems (Perrow, 1984). Machine learning, neural networks, and many other general purpose AI technologies rely on exactly this structure of tightly coupled feedback loops and are highly prone to system accidents

as a result (Maas, 2018; Fourcade & Johns, 2020). System accidents are difficult to understand and challenging to prevent. In systems where no single individual has full knowledge of the system or how its component parts interact, full oversight is impossible, even by highly trained experts (Elish, 2019).

One means of preventing these accidents is to allow sufficient decentralization so that AI workers can factor in localized variables that may impact operations and respond accordingly (Vaughan, 1996). AI workers are often the only individuals with sufficient proximity to these systems to act proactively in calling out potential accidents or harms. Institutions designing AI systems can purposefully make elements of these systems opaque as a means of "self-protection," guarding against industry competition or public scrutiny (Pasquale, 2015; Burrell, 2016). Scholars have demonstrated how dissent channels that draw upon democratic consensus-building can be integrated within AI design, training, and deployment processes (Dobbe et al., 2020). Such mechanisms, however, cannot function properly under the jurisdiction of a single institution, which exist to satisfy their own institutional goals as opposed to greater societal objectives. External arbitrators need to be involved to specify terms and enforce negotiations in a way that can maintain trust and cooperation (Hadfield-Menell & Hadfield, 2019). Collective action has been an important mechanism for AI workers to voice their dissent in the workplace when external arbitrators are not available.

Conclusion

In this chapter, we describe how AI harms are handled within institutions and demonstrate how the positionality of management, ethicists, and workers affect the treatment of the identified harm. Each of these roles comes with its own set of responsibilities for designing and managing AI systems. These institutional roles interact with an individual's existing ethical beliefs, lived experiences, and interests to allow them to come to their own judgment of what constitutes a harm and how best to mitigate this harm.

There are three steps involved in the harm reporting process: identification, the governance decision, and the response. Workers turn to collective action as a response to challenge higher-level decisions when they lack the individual power within their own role. The claims workers make are based on their relationship to the system they are critiquing. Workers may claim they are themselves subject to harm as a result of an AI system. If they played a role in designing or training the system itself, workers may claim control over the product of their own labor or proximate knowledge of the system. Through collective action, AI workers stake a claim for themselves in AI governance in the workplace and assert their perspective as individuals whose labor is entangled with these systems.

Thus far, research in AI governance and labor has focused on the identification stage of the harm reporting process, particularly on providing frameworks by which practitioners can identify and evaluate harms. There has been an open acknowledgment of the ways in which identification is an inherently political process, one that requires questioning existing power structures and inequalities within society. Outlining the two additional steps that follow—the governance decision and the response—allows us to trace the full political

process of harm reporting in the AI workplace and identify where power lies within this process. There is a path dependency in this process that is important to acknowledge. How harms are identified can determine how harms are governed and treated. As research progresses to these second and third steps, it will be necessary to capture and describe these links between identification, governance decision, and response. The model presented here provides a generalizable framework which can enable this research.

Acknowledgments

The authors would like to thank April Mariko Salazar for excellent research assistance. This research was supported by funding from the Jain Family Institute and the Center for Technology, Society, and Policy at the University of California, Berkeley. Earlier versions of this paper benefited from feedback provided by Baobao Zhang, Francis Tseng, Rumman Chowdhury, Lilly Irani, and members of the Jain Family Institute Digital Ethics Reading and Works-in-Progress Group, and participants in the 2021 Genealogies of Data Workshop.

Notes

1. We derive these terms from the processes laid out in Dobbe et al. (2020), although we are aware others have made this distinction or very similar ones.
2. An important distinction must be drawn between AI workers of this category and AI consumers. Consumers use an AI-powered product (e.g., search engine) to fulfill some primary function (e.g., query for information) and provide their behavioral data to companies in conjunction with usage. They are reimbursed solely through access to a product. The "gift economy" business model, whereby consumers receive "free" access to products in exchange for their data, drives many of the world's most profitable companies (Hoofnagle & Whittington, 2014; Fourcade & Kluttz, 2020). Despite the tremendous value they create, we do not consider these consumers in the category of AI workers.
3. Collective action terms include protest*, petition*, strike*, open letter*, walk out*, union*, boycott*, letter*, lawsuit*, discuss*, and negotiat*.
4. Employment terms include employee*, worker*, contract*, and labor*.
5. To be considered "collective," events must involve a minimum of two employees who recognize themselves as a group united by a shared cause and/or issue. The cause and/or issue should be relevant to a broader public, defined as a community which is not directly related to the company through employment or financial ties. The broader public does not include company shareholders, businesses with shared interests, or owners of company property, which include consumers of company products. An action that would typically not be deemed relevant to a broader public on its own merit may qualify if the group presents an argument that it is occurring in response to a cause and/or issue that is relevant to a broader public (e.g., an employee who is believed to have been fired as retaliation for a recent protest).
6. To present "evidence of action," events must involve an attempt to present the cause and/or issue outside of the immediate group. Actions may be either internal (available or visible only to other employees) or external (available or visible to the broader public). Lawsuits may be included if they are granted class action status or if they incite additional collective action. Actions should not be initiated by company management.

7. "Tech workers" are defined as those current or recently employed (within the last year) workers in the technology industry. The technology industry includes but is not limited to information technology, Internet, hardware, software. It does not include adjacent industries (e.g. digital media or the video game industry). It does include online retailers and social media companies. Academics whose research concerns technology and students or interns who are preparing to enter the tech industry can be considered tech workers.
8. Sometimes, harms are identified by the media and then acted upon by workers. The case of the Google Walkout is a good example; a *New York Times* report about the $90-million payout to Andy Rubin catalyzed the Google Walkout (Wakabayashi & Benner, 2018)
9. Harms can be identified by management in theory, although we have not documented such an event in our archive. Research suggests that management is sensitive to reputational risks from AI harms and can be compelled to act to prevent reputational damage (Rakova et al., 2021).
10. Former Google employees, Meredith Whittaker and Claire Stapleton, were retaliated against by Google after publicly voicing criticism of their employer (Tiku, 2019). Amazon fired Emily Cunningham and Maren Costa after their public criticism of Amazon's handling of COVID-19 and poor environmental practices, which included criticism of Amazon's sale of AI capabilities to oil companies (Greene, 2020). Even though the actions taken are collective, the risks and consequences of organizing such actions are often faced by individuals (Amrute, 2021).
11. Ele.me's AI system is named "Ark." According to its engineers, Ark can adjust and optimize deliveries for lunch hours vs regular hours and high rises vs small neighborhoods (Si, 2018). Meituan's equivalent system is called "Super Brain." These systems are allegedly able to help delivery drivers save time by providing them with smarter routes and facilitating a more efficient distribution of orders across the system.
12. China Labor Bulletin recorded 45 strikes by delivery drivers in 2019 (China Labor Bulletin Strike Map, 2020)
13. "We, members of the AI and physical sciences communities, are distressed to find that the intellectual output of our research, which has been performed with the goal of benefiting humanity, will be applied to further deepen the climate crisis. These applications are not only misaligned with our values and moral responsibilities, but they are also in direct conflict with the existential interests of planetary life" (Tech Won't Drill It, 2020). "We have to take responsibility for the impact that our business has on the planet and on people" (Amazon Employees for Climate Justice, 2019).
14. Since its inception, researchers have noted how software is especially error-prone compared to more physical technologies (Parnas et al., 1990). Industry estimates assume approximately 15 to 50 errors per every one thousand lines of non-AI computer code (McConnell, 2004).

References

Abbott, Andrew. (1988). *A system of professions*. University of Chicago Press.

Abend, Gabriel. (2014). *The moral background: An inquiry into the history of business ethics*. Princeton University Press.

Allyn, Bobby. (2020). Google AI team demands ousted Black researcher be rehired and promoted. *National Public Radio*. https://www.npr.org/2020/12/17/947413170/google-ai-team-demands-ousted-black-researcher-be-rehired-and-promoted.

Altman, Micah, Wood, Alexandra, & Vayena, Effy. (2018). A harm-reduction framework for algorithmic fairness. *IEEE Security Privacy 16*(3), 34–45.

Amazon Employees for Climate Justice. (2019). Amazon employees are joining the Global Climate Walkout, 9/20. *Medium*. https://amazonemployees4climatejustice.medium.com/amazon-employees-are-joining-the-global-climate-walkout-9-20-9bfa4cbb1ce3.

Amodei, Dario, Olah, Chris, Steinhardt, Jacob, Christiano, Paul, Schulman, John, & Mané, Dan. (2016). Concrete problems in AI safety. *arXiv preprint* arXiv:1606.06565.

Amrute, Sareeta. (2019). Of techno-ethics and techno-affects. *Feminist Review 123*, 56–73.

Amrute, Sareeta. (2021). A new AI lexion: Dissent. *The New AI Lexicon*. AI Now Institute. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-dissent-2b7861cad5ff.

Belfield, Haydn. (2020). Activism by the AI community: Analysing recent achievements and future prospects. Presentation. AIES '20. Association for Computing Machinery.

Benjamin, Ruha. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Berger, Peter L., & Luckmann, Thomas. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Penguin Books.

Boffey, Philip M. (1986). Software seen as obstacle in developing "Star Wars". *The New York Times*, September 16. https://www.nytimes.com/1986/09/16/science/software-seen-as-obstacle-in-developing-star-wars.html.

Bostrom, Nick, & Yudkowsky, Eliezer. (2011). The ethics of artificial intelligence. In William Ramsey and Keith Frankish (Eds.), in *The Cambridge handbook of artificial intelligence*, pp. 316–334. Cambridge University Press.

Buolamwini, Joy, & Gebru, Timnit. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Presentation. Proceedings of Machine Learning Research. Vol. 81, pp. 77–91. PMLR.

Burrell, Jenna. (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society 3*(1).

Canon, Gabrielle. (2020). Google workers reject company's account of AI researcher's exit as anger grows. *The Guardian*. https://www.theguardian.com/technology/2020/dec/07/timnit-gebru-google-firing-resignation-ai-research.

Caves, Richard E. (2000). *Creative industries: Contracts between art and commerce*. Harvard University Press.

China Labor Bulletin Strike Map. (2020). *China Labor Bulletin*. https://maps.clb.org.hk/?i18n_language=en_US&map=1&startDate=2019-01&endDate=2019-12&eventId=&keyword=&addressId=&parentAddressId=&address=&industry=10308&parentIndustry=&industryName=Food%20delivery.

Crawford, Kate. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dafoe, Allan. (2018). *AI governance: A research agenda*. University of Oxford.

Davenport, Christian. (2009). *Media bias, perspective, and state repression: The black panther party*. Cambridge University Press.

Denton, Emily, Hanna, Alex, Amironesei, Razvan, Smart, Andrew, Nicole, Hilary, & Scheuerman, Morgan Klaus. (2020). Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv*. 2007.07399.

Dobbe, Roel I. J., Gilbert, Thomas Krendl, & Mintz, Yonatan. (2020). Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Presentation. AIES '20. Association for Computing Machinery.

Elish, Madeleine Clare. (2019). Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society 5*.

Employees of Microsoft. (2018). An open letter to Microsoft: Don't bid on the US military's project JEDI. https://medium.com/s/story/an-open-letter-to-microsoft-dont-bid-on-the-us-military-s-project-jedi-7279338b7132.

Eyal, Gil. (2013). For a sociology of expertise: The social origins of the autism epidemic. *American Journal of Sociology 118*(4), 863–907.

Fligstein, Neil. (2001). *The architecture of markets: An economic sociology of twenty-first century capitalist societies*. Princeton University Press.

Fourcade, Marion, & Johns, Fleur. (2020). Loops, ladders and links: The recursivity of social and machine learning. *Theory and Society 49*(5), 803–832.

Fourcade, Marion, & Kluttz, Daniel N. (2020). A Maussian bargain: Accumulation by gift in the digital economy. *Big Data & Society 7*(1).

Friedman, Batya, & Nissenbaum, Helen. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS) 14*(3), 330–347.

Gershgorn, Dave. (2020). 250 Microsoft employees call on CEO to cancel police contracts and support defunding Seattle PD. *One Zero*. https://onezero.medium.com/250-microsoft-employees-call-on-ceo-to-cancel-police-contracts-and-support-defunding-seattle-pd-e89fa5d9e843.

Google. (2021). AI principles: Reviews and operations. *Google*. https://ai.google/responsibilities/review-process/.

Google Walkout for Real Change. (2020). Standing with Dr. Timnit Gebru—#ISupportTimnit #BelieveBlackWomen. *Medium*. https://googlewalkout.medium.com/standing-with-dr-timnit-gebru-isupporttimnit-believeblackwomen-6dadc300d382.

Gray, Mary L., & Siddharth, Suri. (2019). *Ghost work: How to stop silicon valley from building a new global underclass*. Houghton Mifflin Harcourt.

Greene, Daniel, Hoffman, Anna Lauren, & Stark, Luke. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Greene, Jay. (2020). Amazon fires two tech workers who criticized the company's warehouse workplace conditions. *The Washington Post*. https://www.washingtonpost.com/technology/2020/04/13/amazon-workers-fired/.

Haas, Peter M. (1992). Introduction: Epistemic communities and international policy coordination. *International Organization 46*(1), 1–35.

Hadfield-Menell, Dylan, & Hadfield, Gillian K. (2019). Incomplete contracting and AI alignment. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Presentation. AIES '19. Association for Computing Machinery.

Hao, Karen. (2020). We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

Hoofnagle, Chris Jay, & Whittington, Jan. (2014). Free: Accounting for the costs of the internet's most popular price. *UCLA Law Review 61*, 606–669.

Jack, Margaret, & Avle, Seyram. (2021). A feminist geopolitics of technology. *Global Perspectives 2*(1).

Jobin, Anna, Ienca, Marcello, & Vayena, Effy. (2019). Artificial intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence 1*.

Lei, Ya-Wen. (2021). Delivering solidarity: Platform architecture and collective contention in China's platform economy. *American Sociological Review 86*(2), 279–309. https://doi.org/10.1177/0003122420979980.

Lynch, Amanda H., & Veland, Siri. (2018). *Urgency in the anthropocene*. The MIT Press.

Maas, Matthijs M. (2018). Regulating for "Normal AI Accidents": Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. Presentation. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.

McAdam, Doug. (2015). Collective action. In A.D. Smith, X. Hou, J. Stone, R. Dennis, and P. Rizova (Eds.), *The Wiley Blackwell encyclopedia of race, ethnicity, and nationalism*. Wiley. https://doi.org/10.1002/9781118663202.wberen298

McConnell, Steve. (2004). *Code complete: A practical handbook of software construction*, Second Edition. Microsoft Press.

Meyer, John W., & Rowan, Brian. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology 83*(2), 340–363.

Microsoft. (2021). Operationalizing responsible AI. *Microsoft*. https://www.microsoft.com/en-us/ai/our-approach.

Microsoft Employees. (2019). Microsoft workers for climate justice. *GitHub*. https://github.com/MSworkers/for.ClimateAction.

Moss, Emanuel, Watkins, Elizabeth Anne, Singh, Ranjit, Elish, Madeleine Clare, & Metcalf, Jacob. (2021). Assembling accountability: Algorithmic impact assessment for public interest. *Data & Society*. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

Noble, Safiya. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Parnas, David L., van Schouwen, A. John, & Kwan, Shu Po. (1990). Evaluation of safety-critical software. *Commun. ACM 33*(6), 636–648.

Pasquale, Frank. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perrow, Charles. (1984). *Normal accidents: Living with high-risk technologies*. Princeton University Press.

Pichai, Sundar. (2018). AI at Google: Our principles. *Google: The Keyword*. https://blog.google/technology/ai/ai-principles/.

Posada, Julian. (2020). The future of work is here: Toward a comprehensive approach to artificial intelligence and labour. *Ethics of AI in Context*. arXiv:2007.05843.

Posada, Julian. (2021). Unbiased: Why AI needs ethics from below. *The New AI Lexicon*. AI Now Institute.

Raji, Inioluwa Deborah, Smart, Andrew, White, Rebecca N., Mitchell, Margaret, Gebru, Timnit, Hutchinson, Ben, Smith-Loud, Jamila, Theron, Daniel, & Barnes, Parker. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Presentation. FAT* '20. Association for Computing Machinery.

Rakova, Bogdana, Yang, Jingying, Cramer, Henriette, & Chowdhury, Rumman. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact 5*(CSCW1).

Ranganathan, Aruna. (2018). The artisan and his audience: Identification with work and price setting in a handicraft cluster in southern India. *Administrative Science Quarterly 63*(3), 637–667.

Renwu Staff. (2020). Delivery drivers, trapped in the system. *Renwu Magazine*. https://mp.weixin.qq.com/s/Mes1RqIOdp48CMw4pXTwXw.

Rosenblat, Alex. (2018). *Uberland: How algorithms are rewriting the rules of work*. University of California Press.

Rosenblat, Alex, & Stark, Luke. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers. *International Journal of Communication 10*(27).

Russon, Mary-Ann. (2020). Uber sued by drivers over "automated robo-firing". *BBC*. https://www.bbc.co.uk/news/amp/business-54698858.

Selbst, Andrew D. (2017). Disparate impact in big data policing. *Georgia Law Review 52*, 109.

Shane, Scott, Metz, Cade, & Wakabayashi, Daisuke. (2018). How a Pentagon contract became an identity crisis for Google. *The New York Times*, May 30. https://www.nytimes.com/2018/05/30/technology/google-project-maven-pentagon.html.

Shestakofsky, Benjamin, & Kelkar, Shreeharsh. (2020). Making platforms work: Relationship labor and the management of publics. *Theory and Society 49*(5), 863–896.

Si, Ying. (2018). After experiencing five iterations, Ele.me's "Arc" AI dispatching system controls 3 million drivers. *Aliyun Developer Community Blog*. https://developer.aliyun.com/article/594688.

Tan, J. S., & Nedzhvetskaya, Nataliya. (2022). *Collective action in tech archive*. Collective Action in Tech. https://data.collectiveaction.tech/.

Tarnoff, Ben. (2020). The making of the tech worker movement. *LOGIC*.

Tech Won't Drill It. (2020). Tech won't drill it. *Medium*. https://medium.com/@techwontdrillit/tech-wont-drill-it-a63594dc6e66.

Tiku, Nitasha. (2019). Google walkout organizers say they're facing retaliation. *WIRED*. https://www.wired.com/story/google-walkout-organizers-say-theyre-facing-retaliation/.

Useem, Michael. (1993). *Executive defense: Shareholder power and corporate reorganization*. Harvard University Press.

Valiant, Leslie G. (1984). A theory of the learnable. *Communications of the ACM 27*(11), 1134–1142.

Vaughan, Diane. (1996). *The Challenger launch decision: Risky technology, culture and deviance at NASA*. University of Chicago Press.

Wakabayashi, Daisuke, & Benner, Katie. (2018). How Google protected Andy Rubin, the "Father of Android." *New York Times*, October 25. https://www.nytimes.com/2018/10/25/technology/google-sexual-harassment-andy-rubin.html.

Williams, Lauren. (2020). JEDI delay slows DOD's AI push. *Defense Systems*. https://defensesystems.com/articles/2020/05/27/jedi-delay-slows-ai.aspx.

CHAPTER 31

STRUCTURED ACCESS

An Emerging Paradigm for Safe AI Deployment

TOBY SHEVLANE

Introduction

The development of new AI capabilities often brings discussion about whether certain uses of AI should be placed off limits. We see this, for example, with facial recognition technology, where a number of U.S. cities have banned the use of the technology by law enforcement. The companies and research groups developing AI ("AI developers") also have a role to play in shaping how the technology is used. AI developers can engage in AI governance through the software they build and how they choose to deploy that software.

Both the field of AI safety, and the recent initiative of AI "publication norms," provide guidance for how developers can shape the impact of the AI systems they deploy. Research in AI safety, broadly defined, offers insights for building AI systems that have beneficial properties: aligned with human values, responsive to human oversight and control, and difficult to misuse for harmful purposes (Christiano, 2016). On the other hand, the recent initiative of "publication norms for responsible AI" (see, for example, Partnership on AI, 2021) concerns the "publication" phase of AI research and development, where research outputs, including trained AI systems, are disseminated. AI developers are encouraged to exercise caution before sharing AI software (and other research outputs) that might have a harmful effect on the world.

These two projects are much more tightly linked than has hitherto been acknowledged. The utility of building safe AI systems is greatly reduced if those systems are then shared in a manner that allows actors to circumvent the relevant safety features. There are many reasons why actors might rework AI systems to be more capable of causing harm, including malicious intent (Brundage et al., 2018), a negligent disregard for risks (Critch & Krueger, 2020), and structural pressures such as economic or geopolitical competition (Bostrom, 2014; Dafoe, 2015). AI is "dual use" on two levels: an AI system can be directly used for both beneficial and harmful purposes, but additionally, AI systems designed to only serve beneficial purposes can be modified to serve harmful purposes. AI developers, therefore, must

simultaneously address both how their AI systems can be directly used *and* the pathways through which their AI systems can be adapted and built upon.

This chapter introduces an emerging paradigm for AI deployment, which I refer to as *structured access*. Structured access involves constructing, through technical and often bureaucratic means, a controlled interaction between an AI system and its user. The interaction is structured to both (a) prevent the user from using the system in a harmful way, whether intentional or unintentional, and (b) prevent the user from circumventing those restrictions by modifying or reproducing the system.

The deployment of OpenAI's GPT-3 models (Brown et al., 2020) serves as an early example of structured access. Approved users can access the models through a web interface and an application programming interface (API); external developers can build applications that integrate GPT-3 models via the API, subject to approval. Although this arrangement was partly motivated by commercial factors, it was also motivated by a concern that the models could, if open sourced, be used in harmful ways, such as the mass production of disinformation online (OpenAI, 2020). OpenAI's API platform limits the ways in which the GPT-3 models can be used, and developers who want to build applications using the API must conform to certain safety standards. Another example is Google Cloud's Vision API, which has facial recognition capabilities that only work on a particular set of celebrities, with Google limiting who can use this feature.

Both these examples involve models that are stored in the cloud. Alternatively, it is possible to build structured access principles into models that are disseminated and run on users' hardware. This would involve engineering the model to both conform to certain safety standards and be robust against unauthorized modification. However, cloud-based deployment of AI systems is a much more natural fit with structured access. When the AI system is cloud-based, the developer has many more opportunities to put in place mechanisms for shaping how the system is used. These can be built into the AI system itself, but they can also be built into the interface by which users and application developers interact with the system, and the rules governing who can use the system and for what purposes. These protections can also evolve over time, including revoking any capabilities that turned out to be unsafe—something that is much harder if the underlying software has been disseminated. In addition, cloud-based structured access provides greater security against unauthorized modification of AI systems because the user has no local copy of the software to modify.

Most thinking under the "publication norms" banner assumes that AI research outputs (e.g., models, code, descriptions of new methods) will be shared as information and asks whether certain pieces of information should be withheld. Structured access goes one step further. Structured access leverages the fact that AI models are simultaneously both (a) informational (i.e., exiting as files that can be inspected, copied, and shared) and (b) tools that can be applied for practical purposes. This means that there are two fundamental channels through which AI developers can share their work. They can disseminate the software as information (e.g., by uploading models to GitHub). But they can also lend access to the software's capabilities, analogous to how a keycard grants access to certain rooms of a building. Structured access involves giving the user access to the tool, without giving them enough information to create a modified version. This gives the developer (or those regulating the developer) much greater control over how the software is used, going beyond what can be achieved through the selective disclosure of information.

Therefore, this chapter attempts a reframing of the "publication norms" agenda, adding a greater emphasis on arm's length, controlled interactions with AI systems. The central question should not be: *what information should be shared, and what concealed?* (although that is still an important sub-question). Rather, the central question should be: *how should access to this technology be structured?*

Structured access has not yet reached full maturity and, as we will see, there is much room for continued technical and institutional development. Structured access should not be equated with a lack of transparency over the technology: researchers should explore methods for rendering AI systems transparent (in the ways necessary to provide accountability over safety) without leaking too much intellectual property. Nor should structured access be equated with a centralization of power in the hands of technology companies. Decisions over access could be outsourced to other actors or regulated by governments.

I focus on structured access to trained, deep learning models. Nonetheless, the same principles could be extended to cover the software and hardware used to train these models.[1] I do not discuss legal methods for controlling access to AI systems, such as licenses specifying how a model should be used, but these could be used in tandem with the methods discussed.

The rest of the chapter is organized as follows. First, I introduce the concept of structured access. Then, I will draw the distinction between sharing information and granting access to a tool and argues that there are limits to what can be achieved through the selective disclosure of information. Next, I describe how structured access could apply to both locally disseminated software and cloud-based services but argues that the cloud-based option is superior. Finally, I address the criticism that structured access centralizes power in the hands of AI developers.

What Is Structured Access?

Structured access is a paradigm for safe AI deployment. The aim is to prevent an AI system from being used harmfully, whether the harm is intended by the user or not. The developer offers a controlled interaction with the AI system's capabilities, using technical and sometimes bureaucratic methods to limit how the software can be used, modified, and reproduced. The methods of structured access contrast with the distribution of open-source software in that structured access places greater distance between AI capabilities and the actors accessing those capabilities.

Structured access is rooted in a broader phenomenon, going beyond AI, where the owners of potentially harmful artefacts attempt to place limits on how users can interact with those artefacts. For example, certain biological laboratories have the capability to print DNA sequences and offer this as a service. The synthesized DNA can be used for beneficial research but could in theory be used for the creation of bioweapons. This means that the lab must screen orders to ensure that the synthesized DNA will be used safely. In addition, because the printing technology is becoming more widely available, there are calls for DNA synthesis machines to be manufactured with hardware-level "locks" on what they can print (Esvelt, 2018). Both these methods work by adding structure to the interaction between the actor requesting the synthesis and the underlying printing capability.

If structured access means building a controlled interaction between the AI system and the user, then what exactly is being controlled? There are two broad categories: (1) *use controls*, which govern the direct use of the AI system (who, what, when, where, why, how?); and (2) *modification and reproduction controls*, which prevent the user from altering the AI system or building their own version in a way that circumvents the use controls.

Use controls

The purpose of the use controls is to ensure that the AI system is used in a safe and ethical way. The controls govern who can access the AI system and for what purpose, the underlying technical features of the AI system, and the scope of capabilities on offer. Partly this involves careful design of the AI system itself, taking into account the system's safety, its alignment with human values, whether its outputs are fair and unbiased, and whether its capabilities could be misused by actors seeking to intentionally cause harm. These are "software-level use controls." But the use controls need not be built into the AI system itself—they can alternatively be built into the procedures through which the user accesses the AI system. These are "procedural use controls." For example, if a developer wants to avoid their language model being used for medical diagnosis, then rather than, say, removing medicine-related text from the model's training data, the developer could vet users and deny access to anyone hoping to use the model for diagnosis.

Modification and reproduction controls

Let's imagine that the developer has built an AI system to a high standard of safety, putting in place many software-level use controls. The developer then open sources that system, including public sharing of the model's weights, and any code, datasets, and simulated environments used to train the system. The open-source approach benefits researchers and product developers who can now build upon the system, make modifications, and extend the work. Indeed, modifiability is a central tenet of open source software (Kelty, 2008). However, the same modifiability enables the user to remove software-level use controls (e.g., by retraining the model such that it learns a capability that the developer had purposefully avoided). Similarly, the open-source approach makes the AI system reproducible—users can re-run the training regime themselves. This means that they can make changes to the training regime, such as changing the data on which the system was trained or editing the loss function such that it no longer penalizes unsafe behavior. Additionally, even if the developer puts in place procedural use controls on who can access the AI system and for what purpose, if the AI system is easily reproducible, other actors will be able to share the system without these controls.

Therefore, to successfully protect the use controls against removal or circumvention, the developer must also introduce controls on modification and reproduction of the AI system. There are certain obvious, preliminary methods for doing so. The developer can decline to open source the model and the training code. The developer can put in place strong cybersecurity defenses against theft of such information from its servers. However, below we will see more sophisticated modification and reproduction controls. For example, for

models that run on users' hardware, there are methods for making the model immune to fine-tuning. For cloud-based AI services, the developer can go one step further and allow the user to, at arm's length, make authorized (but not unauthorized) modifications to the model.

The Limits of Selective Information Disclosure

Structured access is related to the concept of structured transparency (Bostrom, 2019; Shevlane et al., 2020; Trask et al., 2020). According to Trask et al. (2020, p. 3), "A system exhibits structured transparency to the extent that it conforms with a set of predetermined standards about who should be able to know what, when they should be able to know it, and what they should able to do with this knowledge; in other words, if it enforces a desired information flow." This involves finding mechanisms, both technical and social, for granting access to certain pieces of information while keeping others private (the archetypal example being a sniffer dog, which signals if a bag contains explosives without revealing the other contents of the bag). Structured access differs in that, where structured transparency would govern what information somebody knows and does not know about an AI system, structured access goes further by governing what somebody can and cannot do with an AI system.

There is some overlap between these two. As we saw above, structured access will normally involve limiting access to information about the AI system's design, as part of the controls on modification and reproduction. Blanket secrecy will usually be unnecessary: some forms of information about an AI system will not be very useful for modification and reproduction, and yet will have high social value. For example, for AI safety reasons, developers could grant academic researchers the ability to, at arm's length, run interpretability tests on an otherwise private model, allowing high-level insights to be gleaned about how the model functions. Such opportunities for selective disclosure of information mean that the techniques of structured transparency discussed by Trask et al. (2020) are relevant and useful within structured access.

However, structured access involves more than just controlling the flow of information about AI systems. A file containing the parameters of a deep learning model is not only a piece of information, but also something that, when combined with certain computing infrastructure, can be used as a tool. As philosophers of software have pointed out, software code has a dual nature, as both a piece of text, written in a formal language, which can be read and interpreted, but also as something that functions in a machine-like way (Colburn, 1999). As information, the weights of a language model can be inspected, copied, and shared; but on the right computer, they can also be used for practical purposes, for example, to write the conclusion of an essay, or to classify social media posts into political categories.

Structured access involves separating the software-as-a-tool from the software-as-information, such that the user can access the tool without accessing the information necessary to recreate it.[2] This allows the tool to gain widespread use, even if its design has not been disclosed.[3] It also opens a new set of opportunities for the developer to put in place use controls. We will see many examples of this below; for example, the developer can enforce

rules on what queries are fed into the model; the developer can monitor how the model is being used, shutting down users who violate the rules; or the developer can retract certain capabilities that, with time, appear harmful.

This level of control is not possible if the developer is forced to only rely on selectively disclosing the informational content of AI software. Selective disclosure of information works well for factual information—such as personal medical data or the number of nuclear warheads that a country possesses—where some facts can be communicated and others kept private. However, this does not work so well for the governance of dual use technologies (e.g., for blueprints for the design of a weapon or software code). Here, the information is of a different character: it is "how to" information, with certain practical applications, and those applications are numerous, some beneficial and others harmful. The fundamental problem is that there exists no sweet spot where the user knows enough to achieve only the beneficial ends. Either they know too little, and so cannot use the technology at all, or they know too much, and so they have too much leeway.

This problem is illustrated by DialoGPT, a chatbot built by researchers at Microsoft (Zhang et al., 2019). Due to concerns that the chatbot's outputs were often "unethical, biased or offensive," the researchers open sourced the model but without the final piece of code necessary for producing text from the model. This is no solution to the mixed nature of the model's outputs because either users cannot get the model to produce any text whatsoever, or they find substitute versions of the missing code online and thereby have access to the full functionality of the model. Perhaps users who are determined enough to do the latter are, on average, more likely to use the model responsibly. But even so, this is a very imprecise and insecure method of controlling how people use the model. The lesson is that the developer cannot selectively filter the informational content of the software in a way that neatly discriminates between uses.

On top of this, there are two additional difficulties with using selective disclosure as a means of controlling how people use the software. First, the donor of information cannot later change their mind and revoke the information that they chose to share. This is a serious problem given the difficulty of forecasting the impact of a technology in advance of deployment and given that the impact of the technology might depend on factors that change over time (such as the presence of complementary technologies, or the propensity of actors to misuse the technology). Second, information is easily passed on. If details about an AI system are disseminated to a specific group of recipients, there is no guarantee that they will not disseminate that information further. Different actors will be more or less prone to misusing AI technologies, and so the developer may wish to exercise high levels of control over who can use a particular system. In contrast, if the developer deploys the AI system through an API or web interface, they can require users to log into the service.

Implementation of Structured Access: Local and Cloud-based Systems

AI developers, like software makers more generally, have two broad options for deployment. In one case, the user is given a local copy of the software that they can run on

their own hardware. In the other case, the user has no local copy, and interacts with the model remotely through an interface; the model is stored on an external server. The latter approach reflects the broad trend towards "software as a service" (and related phenomena such as "platform as a service"), whereby today, users commonly interact with software through interfaces such as web browsers (e.g., Facebook, Google Docs), without necessarily downloading the software onto their computers.[4]

AI developers currently rely on both these modes of deployment. It is common for research groups, including those within large AI companies, to publicly share copies of their deep learning models on sites such as GitHub, where the models are free to download. Another example is that, where AI software is used on smartphones (such as augmented reality filters for videos), a local copy of the software will often be stored on the phone. On the other hand, AI software is also widely built into the backend of cloud services, such as Google Search or Grammarly (which uses deep learning models stored in the cloud to correct users' grammar). Some AI developers also bundle API model access together with their cloud computing services.

Structured access can be implemented on both local and cloud-based systems. However, we will see that cloud deployment is a better home for structured access, offering greater opportunities for controlling what the user does with the AI system. The cloud-based approach should be viewed as the archetypal way of implementing structured access.

Local deployment

Under the local deployment approach to structured access, the AI system runs on the user's hardware. The AI system will contain software-level use controls, as well as software-level protections against unauthorized modification and reproduction (such as obfuscation of the model). The result of these controls will be that, although the model sits on the user's hardware, the user has only a constrained interaction with the model.

Use controls

Where the developer distributes local copies of AI software, the obvious place for use controls is at the software level. In the case of deep learning models, this will often involve careful engineering of the model itself. Alternatively, because AI systems often combine multiple deep learning models and other algorithms, the safety mechanism could be more modular; for example, a language model for text generation could be used alongside a model that classifies outputs and filters out certain kinds of statements.[5]

One approach is to construct the training regime such that certain dangerous capabilities are never learned by the model. A rudimentary example occurred with OpenAI's GPT-2 models. The developers trained a range of model sizes, initially only sharing smaller versions, believing these smaller models to be less capable of misuse (e.g., producing less convincing fake news articles) (Solaiman et al., 2019). However, in this case, it was not only the risky capabilities that were weaker—the smaller models were generally less performant. Indeed, manipulating model size is a very imprecise method for controlling how the model is used. Because of this, as well as the lack of controls on modification, GPT-2 should not be considered an example of structured access—it is an example of selective disclosure.

The developer can also put in place procedural use controls, although the scope for doing so is greatly reduced in the case of local deployment. For example, instead of publicly sharing the software, the developer could decide to disseminate the software to only a select group of users. Moreover, the developer could use a software license key system, where the software is only usable once a valid key has been entered. Such a system could in theory help to limit the range of users who can access the software, and even allow the developer to revoke access to the software (if the license is time-limited). However, these methods are likely to be less secure than the equivalent in cloud-based deployment. In addition, unlike the cloud-based approach, there is no way for the developer to directly monitor and police how the software is being used.

Modification and reproduction controls

If the user has a local copy of the model, one threat is that the user will make modifications to the model, for example, by fine-tuning on a new task, or going into the code and removing features that were installed for safety reasons. Another threat is that the user gleans reproduction-relevant information from the model. This could be done through inspecting its architecture or copying elements of the model, such as the weights of certain layers of the network. Alternatively, the user could train a new model using the outputs of the existing model as data; that is, model stealing (Tramèr et al., 2016).

A small body of literature offers technical methods for securing deep learning systems against unauthorized modification and reproduction. The main themes of the literature are preventing fine-tuning and making the neural network less transparent (e.g., via obfuscation or encryption). Three methods are outlined in Table 31.1.

Table 31.1 Technical innovations in the protection of deep learning models against modification and reproduction.

Method	Description
"Deepobfuscation" via knowledge distillation (Xu et al., 2018)	The model is distilled (i.e., a new model is trained on its outputs) into one that is smaller and has a simpler structure. This conceals the structure of the initial model; and the new, smaller model less well-suited to fine-tuning.
"MimosaNet": Increased sensitivity to weight changes (Szentannai et al., 2019).	Fine-tuning is prevented by making the model "extremely sensitive" to changes in the model weights. This is achieved through adding additional weights to the model which are designed to leave the performance of the model unchanged, yet which dramatically alter the performance of the model if they are modified.
Encryption of the model weights (Alam et al., 2020; Chakraborty et al., 2020; Xue et al., 2021)	The model is trained such that the input of a secret key is necessary for it to perform well. The secret key can also be stored on trusted hardware. This allows the developer to limit who can use the model. In addition, the encrypted weights are resistant to fine-tuning.

These methods are specific to deep learning models. Generally applicable methods of protecting software against piracy are also relevant. For example, one technique would be to compile the model into binary form before sharing it with users. (Nonetheless, there is no guarantee that a well-motivated attacker could not recover the model, especially if tools for decompiling become more sophisticated.) Another approach is to embed the software within hardware that is difficult for the user to access, as can be done with autonomous vehicles.

Generally, applying structured access to locally deployed AI systems is an uphill struggle because it is harder to control how the user interacts with software that is running on their own computer. Nonetheless, developers may sometimes be forced to rely on the local deployment approach where the cloud-based approach is unavailable. For example, in the case of autonomous vehicles, the system cannot rely on always having a strong internet connection. Such cases motivate further technical innovation in structured access for local deployment.

Cloud-based deployment

The second implementation is where the user never gains possession of the model, but instead interacts with the model through an interface. This is the strategy adopted by OpenAI for their GPT-3 models, where users are given an API and a web interface. Strictly speaking, this category of deployment approaches need not always involve a cloud service. For example, DeepMind initially disclosed to biologists predictions about the structure of COVID-19 from their AlphaFold 2 model (Jumper et al., 2020),[6] without building a cloud interface for biologists to use. The essence is that the model is stored and run on hardware that is not in the possession of the users.

Use controls

As with local deployment, there is ample room for software-level use controls. The developer could then grant the user access to the full functionality of the model through an interface. However, cloud-based deployment excels in the area of procedural use controls.

Instead of having to encode distinctions between different uses into the software itself, the developer can set enforceable rules about what users can do with the model. For example, with GPT-3, OpenAI has different categories of uses: those which are considered generally safe (e.g., extracting keywords from a passage); those which are disallowed (e.g., using GPT-3 for spam); and those which will be evaluated on a case-by-case basis, depending on the sensitivity of the domain and other considerations (e.g., using GPT-3 for social media posts). Compliance with such rules can be enforced because the developer can monitor how users are interacting with the model and block access from users who violate the rules. (Monitoring can be done in a privacy-preserving way; see Trask et al., 2020.) The developer can also change the rules over time.

Generally, the developer can create a highly mediated interaction between the model and the user, with multiple layers of use controls in place. As well as rules that separate between different kinds of uses, the developer can also separate between different kinds of user. For

example, the developer could only grant access to actors who are unlikely to misuse the model or could offer tiered access for different types of actors. The developer can also impose rate limits or quotas on the number of times that each user can query the model.

There may also be a layer where external developers build applications on top of the model, with end users interacting with those applications. With OpenAI's GPT-3 models, this arrangement helps with refining the range of possible uses of GPT-3 because the applications are bounded in scope. For example, OpenAI is keen to avoid downstream developers from building an open-ended text generator or chatbot using GPT-3, whereas their documentation clarifies that a chatbot constrained to only answering mathematics questions would likely be acceptable (OpenAI, 2021).

An interesting development is that, with GPT-3, the application developer can use "prompt engineering" to achieve this narrow focus. For example, the application's code could insert the user's question into a standardized template submitted to the API, such as the following:

QUESTION: What is the term for the long side of a triangle?
ANSWER: The hypotenuse.
QUESTION: What is the mathematical term for a whole number?
ANSWER: An integer.
QUESTION: Who was the Democratic candidate in the 1996 U.S. presidential election?
ANSWER: Sorry, I only answer questions about mathematics.
QUESTION: [**Insert user's question**]
ANSWER:

GPT-3's continuation can then be submitted to the user as the response from the question-answering chatbot. The hope is that the model will generalize from the examples given in the prompt and only answer questions about mathematics. The application developers can experiment with different prompt templates to find one that produces the desired kinds of answers. Moreover, the API provider can monitor what prompts are being fed into the API, and so in theory could notice if the application developers try to make the chatbot more open-ended than was previously promised. This is a neat illustration of how AI developers, using cloud-based structured access, are able to implement a more holistic approach to safety, manipulating not only the model's internal makeup but also the process through which the user accesses the model's capabilities.

Modification and reproduction controls

In the extreme case, the user has no access to the inner workings of the model. The user therefore cannot directly modify the model, nor inspect it for the purposes of recreating a modified version. In addition, the developer can install checks against model stealing (i.e., where users train a new model by using the original model's outputs as data), such as by imposing quotas on the number of times the user can query the model.

Nonetheless, the cloud-based approach is sufficiently flexible to accommodate controlled modification and inspection of the model. For example, at the time of writing, OpenAI plans to allow for fine-tuning of its GPT-3 model. Modification could be done at arm's length, with the developer vetting proposed modifications on safety and ethical grounds.

Developers could also build functionality for external researchers to subject the model to interpretability analyses, as a way of facilitating study of the model's safety. An interesting question is how much depth of analysis could be permitted without giving away too much knowledge relevant for recreating the model (e.g., could researchers study specific neurons, or only the more high-level functioning of the model, such as the activations of the different layers?). One arrangement for preserving the privacy of the model could be that the external researcher uploads to the developer a tool that automates the interpretability analysis, then the developer runs that tool on the model and sends back the results. Another option is that the cloud-based interface grants deeper model access to approved, trustworthy researchers, who agree not to share reproduction-relevant information about the model. Overall, improvements in the functionality of these interfaces should allow cloud-based structured access to recoup some of the advantages of open source software (i.e., modifiability and interpretability).

Finally, one problem with the cloud-based approach is that it currently requires the developer to invest in the supporting infrastructure, which might be too burdensome for lower-resource developers of AI systems, including many academic labs. One possibility to be explored is the creation of an organization that acts as a centralized hub for cloud-based structured access; that is, the structured access equivalent of GitHub or Hugging Face. This platform could host many different models from many different developers and employ experts to make standardized and externally informed decisions about what capabilities should be accessible to whom.

The Centralization of Power

An obvious criticism of the vision of structured access presented so far is that it centralizes power in the hands of AI developers. One concern might be that the developers cannot be trusted to make the right decisions about what capabilities should be granted to whom, especially if they have a commercial interest in selling AI services. A related concern is the possible lack of accountability in how AI developers make these decisions.

The first step to mitigating these concerns is acknowledging that structured access can be embedded in a wider governance framework. For example, when governments regulate the use of AI, they could require AI developers to avoid undermining those regulations when providing access to AI systems. This would be analogous to the regulation of underage drinking, where governments ban retailers from selling alcohol to underage individuals. In addition, governments could regulate AI deployment in a more hands-on way, setting standards for what capabilities should be made available, in what form, to whom. In tandem, governments could require that AI developers facilitate some minimum level of access by external researchers who perform interpretability tests on the relevant AI system. Alternatively, these standards could be set by independent civil society bodies.

We can also cast doubt on the pervasive assumption that says widely distributing AI knowledge will, accordingly, widely distribute power. There are two reasons to doubt this. First, if users gain unencumbered access to the technology, what has been distributed? The power to set communal standards over use of the technology has not been distributed to users. In Hohfeldian terms, the users are each granted a *freedom*; that is, a freedom to use

the technology how they like, rather than a "power" over how others use the technology (see Hohfeld, 1919). The only way to actually distribute the power that structured access (de facto) gives to the developer is to embed structured access in a wider governance framework. Another way of putting the same point: there are two different definitions of AI "democratization."[7] One definition is simply that a wide range of people can use the technology. The other definition is more faithful to the meaning of "democracy" in the political arena and asks whether users can collectively influence how the technology is designed and deployed. These two types of democratization sometimes conflict: if AI developers open source their systems, this grants wide access to the system, but simultaneously undermines attempts to build communal decision-making structures within AI deployment.

Second, we can ask: what are the downstream consequences, in terms of the distribution of power in society, of all actors gaining unencumbered access to AI technologies? For many AI technologies, there are reasons to doubt that actors with low levels of existing power would be the primary beneficiaries. As a crude analogy, consider what happens when modern handcuffs become widely available. Handcuffs assist law enforcement officers in making arrests because the officers are legally permitted to use them—it is not that the power to make arrests becomes more evenly distributed. Likewise, there are many AI technologies that will likely increase the power of actors with existing power resources, such as legal authority, data collection abilities, and financial resources (see Pasquale, 2016). For example, an employee cannot come into work one morning and ask to use an AI-powered worker surveillance system on their boss. This is a complex issue, and deserves much more in-depth discussion (see Shevlane & Dafoe, 2021). But suffice to say, there is no simple relationship that links wider availability of AI systems with an equalization of power in society. In fact, structured access could be used for preventing certain actors from abusing their power—for example, by denying access to governments who wish to use AI for political repression.

Conclusion

The theory presented in this chapter should be integrated into an updated understanding of the "publication norms" agenda within AI governance. The publication norms agenda is defined by the Partnership on AI, for example, as "the consideration of *when and how to publish novel research* in a way that maximizes benefits while mitigating potential harms" (Partnership on AI, 2021, emphasis added). In discussions about publication norms, the focus is often on the developer's ability to make certain forms of information public (i.e., research outputs, and discussions about risk), whilst potentially carving out certain research outputs to be kept private. Going forward, AI governance must focus on *both* the sharing (and withholding) of information about AI systems *and* the growing infrastructure for hosting arm's length interactions between users and AI systems.

The early examples of structured access are encouraging, but there is much room for improvement. Two areas should be the focus. First, there is the functionality of cloud-based interfaces. As I argued above, these interfaces could grant users and researchers greater access to the model whilst still maintaining robust controls on modification and reproduction. The user could be given the ability to fine-tune (or otherwise modify) the model in a

controlled setting. External researchers could also be given the ability to run interpretability tools on the model. There is also the question of user privacy: how can the privacy of the user's activity be respected, even though the use of the model is being monitored for compliance with safety standards? Generally, if the interface's functionality can be improved without sacrificing on safety, it is important that these improvements are implemented, such that cloud-based structured access can recover some of the benefits of the open-source approach.

Second, there is the overarching framework of governance within which these cloud-based interfaces sit. One possibility, which can be described as "GitHub for structured access," would involve creating a central repository for models. This repository would be overseen by a body that makes standardized, informed decisions about how access to each model should be structured, while also monitoring compliance with these standards. This would benefit research groups and AI companies who do not have the internal capacity for implementing structured access. Another possibility is that structured access is given legal backing, with governments setting regulatory standards for AI deployment. This could increase the legitimacy and uniformity of decisions over access to models. Overall, structured access is in its infancy, and there is much room for technical and institutional development.

Acknowledgments

I am grateful for helpful comments and suggestions from: Markus Anderljung, Carolyn Ashurst, Ondrej Bajgar, Avital Balwit, Nick Bostrom, Justin Bullock, Allan Dafoe, Eric Drexler, Ben Garfinkel, David Krueger, Jeffry Ladish, Alex Lintz, Luke Muehlhauser, Laurin Weissinger, and Baobao Zhang.

Notes

1. For example, when developers buy cloud compute for training a model, the provider could check that the proposed model conforms with certain safety standards. Brundage et al. (2020, p. 32): "Cloud providers already employ a variety of mechanisms to minimize risks of misuse on their platforms, including "Know Your Customer" services and Acceptable Use Policies. These mechanisms could be extended to cover AI misuse. Additional mechanisms could be developed such as a forum where cloud providers can share best-practices about detecting and responding to misuse and abuse of AI through their services."
2. Note that it is often impossible to completely isolate the software-as-a-tool such that no information about the workings of the tool leaks out during use. If an alien came to earth and encountered our cars for the first time, they could glean some insights about how the car works simply by driving one around, without needing to see any blueprints. In the same way, somebody interacting with an AI system through a web interface could pick up on certain reproduction-relevant information. One simple insight is that the AI capabilities on display are actually achievable. They could also consider the form that the inputs to the model must take (for example, GPT-3 is an autoregressive language model, and, as a consequence, it will predict how a passage of text might be continued, rather than filling in words in the middle of a passage).

3. This vision of AI deployment is highly compatible with Drexler's (2019) vision of "comprehensive AI services," where advanced AI systems are deployed in service of bounded tasks across the economy.
4. Note that many services, especially smartphone apps, rely on a mixture of locally downloaded software and cloud-based software.
5. That said, modular features might be easier to remove through modification.
6. Jumper et al. (2020): "We recently shared our results with several colleagues at the Francis Crick Institute in the UK, including structural biologists and virologists, who encouraged us to release our structures to the general scientific community now."
7. See Divya Siddarth, on the *Radical AI Podcast*, May 27, 2021: "I think the word democratization has come to mean increasing access to a system as being democratizing that system. . . . And I think instead thinking about democracy as the structure of power relations and as a way to provide meaningful voice and agency over decision making rather than simply access is really crucial."

References

Alam, M., Saha, S., Mukhopadhyay, D., & Kundu, S. (2020). *Deep-lock: Secure authorization for deep neural networks*. *arXiv:2008.05966 [cs.LG]*.

Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford University Press. http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=1743975.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy 10* (4), 455–476. https://doi.org/10.1111/1758-5899.12718.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *ArXiv:2005.14165 [Cs.CL]*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv:1802.07228 [Cs.AI]*.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv:2004.07213 [cs.CY]*.

Chakraborty, A., Mondal, A., & Srivastava, A. (2020). Hardware-assisted intellectual property protection of deep learning models. *Proceedings of the 57th ACM/EDAC/IEEE Design Automation Conference*.

Christiano, P. (2016). AI "safety" vs "control" vs "alignment". *AI Alignment*. https://ai-alignment.com/ai-safety-vs-control-vs-alignment-2a4b42a863cc.

Colburn, T. R. (1999). Software, abstraction, and ontology. *The Monist 82* (1), 3–19. https://doi.org/10.5840/monist19998215.

Critch, A., & Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). *arXiv:2006.04948 [cs.CY]*.

Dafoe, A. (2015). On technological determinism: A typology, scope conditions, and a mechanism. *Science, Technology, & Human Values 40* (6), 1047–1076. https://doi.org/10.1177/0162243915579283.

Drexler, K. E. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence (Technical Report #2019-1). Future of Humanity Institute, University of Oxford.

Esvelt, K. M. (2018). Inoculating science against potential pandemics and information hazards. *PLoS Pathogens 14* (10). https://doi.org/10.1371/journal.ppat.1007286.

Hohfeld, W. N. (1919). *Fundamental legal conceptions* (W. W. Cook, Ed.). Yale University Press.

Jumper, J., Tunyasuvunakool, K., Kohli, P., Hassabis, D., & The AlphaFold Team. (2020, August 4). *Computational predictions of protein structures associated with COVID-19*. Deepmind. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

Kelty, C. M. (2008). *Two bits: The cultural significance of free software*. Duke University Press.

OpenAI. (2020, June 11). *OpenAI API*. OpenAI blog. https://openai.com/blog/openai-api/.

OpenAI. (2021). OpenAI API documentation. https://beta.openai.com/docs/api-reference/introduction.

Partnership on AI. (2021). *Publication norms for responsible AI*. The Partnership on AI. https://www.partnershiponai.org/case-study/publication-norms/.

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Shevlane, T., Garfinkel, B., & Dafoe, A. (2020). Contact tracing apps can help stop coronavirus. But they can hurt privacy. *The Washington Post*. https://www.washingtonpost.com/politics/2020/04/28/contact-tracing-apps-can-help-stop-coronavirus-they-can-hurt-privacy/.

Shevlane, T., & Dafoe, A. (2021). AI as a social control technology. Draft manuscript.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release strategies and the social impacts of language models. *ArXiv:1908.09203 [Cs.CL]*.

Szentannai, K., Al-Afandi, J., & Horváth, A. (2019). MimosaNet: An unrobust neural network preventing model stealing. *arXiv:1907.01650 [cs.LG]*.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *ArXiv:1609.02943 [Cs, Stat]*. http://arxiv.org/abs/1609.02943.

Trask, A., Bluemke, E., Garfinkel, B., Cuervas-Mons, C. G., & Dafoe, A. (2020). Beyond privacy trade-offs with structured transparency. *arXiv:2012.08347 [cs.CR]*.

Xu, H., Su, Y., Zhao, Z., Zhou, Y., Lyu, M. R., & King, I. (2018). DeepObfuscation: Securing the structure of convolutional neural networks via knowledge distillation. *arXiv:1806.10313 [cs.CR]*.

Xue, M., Wu, Z., Wang, J., Zhang, Y., & Liu, W. (2021). AdvParams: An active DNN intellectual property protection technique via adversarial perturbation based parameter encryption. *arXiv:2105.13697 [cs.CR]*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2019). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *ArXiv:1911.00536 [Cs]*. http://arxiv.org/abs/1911.00536.

CHAPTER 32

AI, COMPLEXITY, AND REGULATION

LAURIN B. WEISSINGER

Introduction

Artificial intelligence and machine learning (AI) are increasingly commonplace. Some applications of AI, like medical imaging analysis, produce extremely useful results when used by experts and increasingly without human intervention (Pesapane et al., 2018; Alexander et al., 2020). However, many other AI systems do not work as expected or advertised (Narayanan, 2019). For example, the use of AI in hiring, content moderation, and copyright filtering (Gillespie, 2020; Llansó et al., 2020, p. 3), often do not live up to goals and manufacturer promises. Other skeptics, like Manheim and Kaplan (2019), note the destructive potential of AI overall. At the same time, there is unmet demand for specialists with relevant development and management skill sets (Anton et al., 2020).

As earlier chapters of this text have explained, issues with current AI systems are already manifold and include bias (Röösli et al., 2021; Ferrer et al., 2020), lack of transparency (Larsson & Heintz, 2020; Fine Licht & Fine Licht, 2020), lack of fairness (Zhang et al., 2020), and errors not corrected by human oversight (Banerjee & Chanda, 2020). Due to the increasing effectiveness and use of AI, but also the growing concerns about ineffectiveness and harm, public calls for AI regulation, including outright banning certain uses like facial recognition, are getting louder and more numerous (Pasquale & Malgieri, 2021).

This chapter discusses why regulating AI is a challenge but technically feasible if there is political will to do so. First, it will discuss the issue of complexity and other factors that complicate AI regulation, arguing that regulating AI will be difficult, imperfect, and resource intensive. Second, this chapter discusses harms and potential ways to reduce them to then discuss how regulation should be designed. This chapter concludes that complexity is an obstacle to regulation. However, power imbalances, political economy, and lack of political inclination are likely the more relevant concerns.

AI and Complexity

The complexity of AI systems in their technical, social, political, and organizational contexts is high and increasing. This complexity is an obstacle for understanding, designing, and building technical as well as regulatory systems, especially if they are to meet inherently complex objectives like fairness or transparency.

The term complexity is used in a variety of ways in different scientific and technical fields. In this chapter, complexity shall be used in its basic, general meaning: a complex system or entity consisting of many different interacting aspects, parts, or functions, which together form a whole that has different qualities than its parts, and which cannot be described, analyzed, or predicted easily because of its intricacy and internal interactions.

Complexity is an issue not specific to AI but a consequence of our organic (Durkheim, 1984), technologically advanced society (Rescher, 2020). Our potential impact on the social and natural environment has increased alongside ever greater specialization, forcing us to address seemingly endless risk (Beck, 1986; Beck & Beck-Gernsheim, 1994). With higher complexity come greater unknows; some of these are "known unknows," meaning that there is some understanding of the problem and potential outcomes. Complexity also gives rise to "unknown unknowns," or challenges whose existence we are not aware of (Pawson et al, 2011). Both matter to regulation and policy making. The former are the aspects regulation can address relatively easily—at least in theory —while the latter can be mitigated, but not avoided.

Complexity pertains to both known and unknown unknowns: lack of full understanding increases the resources, time, and effort needed to design and build systems that avoid, or at least reduce, negative side effects and externalities. Furthermore, complexity can create challenges outside expected parameters; these can be mitigated—but mitigation is always probabilistic and usually resource intensive.

AI complexity is not only technical but also social and organizational: what we call AI or AI systems are not, at least in practice, singular devices or programs. Instead, they are socio-technical networks of subsystems. This is not uncommon in contemporary value chains (Weissinger, 2020) but in the case of AI and computing, an individual subsystem is already uncertain and carries the risk of emergent and unexpected behavior.

When building contemporary AI systems, multiple types of expertise, components, and subsystems are required. These subsystems form networks of dependencies rather than a hierarchical or tree structure; to function properly, a software component not only requires hardware and operating systems (i.e., a technology stack) but also various APIs, packages, and external services that are often created and serviced by different parties. Networked systems and the resulting complexity are thus one step up from stack complexity; that is, where complex components are built on top of other complex components.[1] Some parts of such larger systems are "building blocks" that will not or only slightly change, but other aspects or subsystems will vary case-by-case, per organization, per sector, or overtime.

While in production use, AI systems continue to be networked: usually, there is not one single machine, algorithm, or data source but different, more or less independent subsystems that contribute functionality, data, or Input/Output more generally. With the rise of the "Internet of Things" and edge computing, AI will increasingly interact directly

with other technological systems, including other AI, creating diverse, sizable, and hardly graspable networks of interaction and interdependence. The more nodes are added to a (systemic) network, the more the universe of possibilities grows, increasing uncertainty and the potential for emergent behavior. Small aberrations, changes, or errors can lead to global effects (Schelling, 2006). In IT, various approaches exist to manage large complex networks and systems, ranging from operations centers, DevOps, DevSecOps, and Site Reliability Engineering (SRE) practices (Beyer et al., 2016). While imperfect and usually resource-intensive to sustain, these tools have contributed to better maintenance and management of complex IT systems.

Socially, various expert providers will continue to emerge (Hirsch et al., 2009) to build and service AI systems. Thus, organizations using, administering, or hosting AI will have limited visibility, with no one, including programmers and designers (Magnusson Sjöberg, 2021), having a full grasp of the systems they use or support. Under these circumstances, allocating liability and responsibility can be difficult, as it might be the very combination of systems that causes failure or harm.

Even narrow AI systems—as they are predominantly used today—often rely on sophisticated algorithms and software packages from different vendors, optimized platforms, plus various internal and external data sources. These systems can and are being audited, sometimes according to set frameworks (Saleiro et al., 2019). With increasing complexity, such audits will become more difficult and expensive, particularly if they go beyond the technical aspects. Furthermore, efficacy is specific to circumstances. If a system shows certain properties in one situation, outcomes might be very different and unpredictable should conditions change. Even if change happens at a micro-level, global consequences are possible (Schelling, 2006). As Michel (2021) observes for military AI systems, differences between testing environments and the real world are particularly pronounced.

Due to these developments, questions relevant for AI regulation move from "smaller worlds," where alternatives and probabilities are usually clearer, toward "larger worlds," dominated by a lack of information and increased uncertainty (Gigerenzer & Gaissmaier, 2011). The "networked complexity" (Weissinger, 2020) layer, in particular, is leading to a much "larger" world (Gigerenzer & Gaissmaier, 2011): the number of relevant, interconnected systems and ties—and thus possibilities for error—are exponentially greater than when we are focusing on "mere" technology stacks. Emergence and systemic evolution (Goldstein, 2011), key features of complex systems, also increase in likelihood and extent, alongside greater opportunities for adversarial action and unexpected errors.

Therefore, AI regulation can never be perfect; not only will it be impossible to address unknown unknowns, but it will also be difficult to deal with known problems. Pawson et al. (2011) argue that "learning over time" might be the only possibility to truly improve mastery over complex systems. While that observation is likely correct, it is also possible to prepare for this complexity challenge: for example, critical systems can be identified, designed with greater safeguards, and can avoid high-uncertainty architectures, technologies, and dependencies. By studying past and potential failure scenarios, critical junctures or possible exception points can be identified and treated. All these measures are imperfect and contingent but do reduce risk.

Crucially, the results of the use of AI we observe are not just the results of a specific algorithm or technology but are more so the results of economic, political, and organizational choices, including the appetite for or the acceptance of risk and possible (externalized)

harm. For example, while Amazon's use of algorithms to track workers is debatable, it might be less the technology at fault and more how it is used. Instead of using AI and technology to optimize processes and identify opportunities for improvement, it appears to be Amazon's choice to use a known limited, superficial, incomplete system to decide who gets fired. According to Soper (2021) Amazon also chose not to provide effective human moderation, penalizing and terminating workers for issues outside their control, like long lines at Amazon's own warehouses.

Harms, Objectives, and Complexity

Regulation is often used to address harms and risks by shaping behaviors or systems. Thus, to understand how to regulate AI, we must first grasp the many open questions surrounding the space and its potential harms. To do this, we can categorize many of the discourses around AI harms and issue spaces into four, overlapping groups related to complexity:

1. Technical vulnerabilities and errors in AI systems
2. Issues or harms that emerge through, or are created by, the use of AI
3. Issues or harms that are magnified considerably by the use of AI
4. Mismatches between AI and social expectations or empirical realities

Technical issues describe technical system errors, where AI systems can be misled by circumstance or adversarial action. For example, Povolny and Trivedi (2020) were able to mislead instances of the Tesla autopilot, "tricking" the system into acting wrongly by slightly changing optical cues. These vulnerabilities are like computing security concerns: inputs are not being parsed correctly, or functions fail to execute. Complexity increases the risk of incidents: the more data are being created and processed, and the more systems interact, the more likely errors are to arise. Furthermore, complexity also decreases our ability to diagnose, understand, and fix problems. For example, unlike expert systems, contemporary stochastic systems cannot be perfectly reverse engineered (Bullock, 2020, p. 494), denying our understanding of individual decisions and the overall logic.

AI can create challenges that did not exist previously. Besides issues of consent (Selinger & Hartzog, 2020), facial recognition technologies are so effective (Tolba et al., 2008; Kaur et al., 2020) that AI adds a whole new concern to video surveillance. Systems can now identify and individually track people in real time. Such levels of automated, untargeted surveillance were previously impossible. Complexity matters to this kind of problem because we do not have sufficient insight into system capabilities, owners, use cases, locations, and interactions.

AI systems often magnify already existing harms. Biases in hiring[2] are well known but can be made worse by AI, if only by the fact that machine outputs are considered more "objective" (Barocas & Selbst, 2016; Sánchez-Monedero et al., 2020; Skitka et al., 1999). Complexity also affects this aspect: organizational, social, and technical systems, each with their own biases and issues, are interacting and confounding outcomes and complicating audits.

Certain areas or spaces are difficult to navigate with AI systems, even if they work flawlessly. Health care, mental health, and the justice system are considered intrinsically "human," benefiting immensely from, or clearly requiring, human contact (Wachter et al., 2020). For example, people have considerable reservations about "robot judges," even though they might be able to compensate for some human failings (Chen et al., 2021). Complexity matters to this aspect because it muddles discussions of scale, impact, and appropriateness. As an example, recidivism is not, and cannot be, (easily) measured (Dressel & Farid, 2018), with rearrest being used as a proxy. However, discourse and expectations revolve around recidivism, not rearrest. Such categorical mistakes are not at all limited to AI but linked to complexity: even fewer people understand AI systems than, say, statistical analysis (Roten, 2006; Gardenier & Resnik, 2002).

Socio-economic challenges

The power imbalance of real-world use is a further concern not directly related to complexity. AI is predominantly used by the powerful and well-resourced on those with less power and fewer resources. Mismatches between what an AI system does and what decision-makers and employees (are made to) believe are common. However, even if these weaknesses are known, AI users like Amazon (Soper, 2021) might continue screening workers—who have little to no recourse—with flawed systems.

The complexity of AI systems coupled with their (related) lack of transparency can hide *administrative evil*. Not only is there a risk of *goal displacement* where *organizations pursue intermediate and easily measurable goals* (Young et al., 2019, p. 8) over actual understanding and real-world improvements. But also, the deployment of such automated systems and approaches risks harm to individuals, particularly those belonging to marginalized communities, by reducing agency and increasing rigid, top-down control.

In addition to furthering imbalance between the weak and powerful while creating a veil of rationality and scientificity, AI systems are often deployed and used in a discriminatory manner (Williams 2020). AI-based false arrests and incarceration unsurprisingly fall among racial lines: multiple black men have been arrested based on erroneous outputs from automated facial recognition systems (Hill, 2020b; Hill, 2020a), apparently with very little fact checking on the side of law enforcement. Thus, while the technology of AI and associated issues of complexity are a challenge, the key harms are created by how AI is used in society. Automated systems are often used as certain institutions "see fit"; for example, the human review supports the customer to get the results they want. Yet, similar activities are not undertaken when it comes to those being acted upon, as demonstrated by the case of *ShotSpotter* (Ropek, 2021). The perceived objectivity of (such) imperfect AI systems may further contribute to harms, if complex, obscure, and indeterminate (social) problems are being resolved through automation (Young et al., 2019, p. 11). The effects of replacing human actors with machines are yet to be uncovered in detail (Bullock, 2020, p. 492); thus, we are even further away from managing these transitions and its (potential for) harm.

While specific issues like facial recognition are closely related to AI, societal questions like corporate and state mass surveillance (Zuboff, 2019; Foucault, 2012), systemic, often inherently complex and multifaceted, social inequalities like racism, xenophobia (Lee, 2021),

caste systems (Agrawal, 2014), social class (Woodward et al., 2021; Lahtinen et al., 2020), and sexism (SteelFisher et al., 2019) are too fundamental to be resolved by specific regulation.

AI regulation could potentially ensure that AI does not make existing problems worse but cannot resolve the wider social issues. The above set of global challenges are inherently complex. They lead to multifaceted, intersecting, and thus differing individual challenges on the ground, which are ideally addressed case-by-case (Bullock, 2020). Society is one of the most complex systems of all, and resolving such conflicts will be difficult due to power imbalances, hardened beliefs, micro–macro interactions (Schelling, 2006), and many more confounding factors.

Complexity of harms

With AI systems becoming more commonplace and complex, harms can grow more difficult to understand and diagnose. Representative amounts of data are generally inaccessible for independent third parties, as well as those affected by the outputs of AI systems. Furthermore, a simple algorithm drawing from a limited number of variables is easier to diagnose and adjust than highly intricate systems relying on disparate inputs from a multitude of data sources and other, often external, systems. Networked systems, data, inputs, and processing are often controlled by different entities and components change independently over time. This further complicates the analysis of how or why certain outputs came to be. Thus, when it comes to individual decisions (e.g., a loan being denied or bail not being granted), it is resource-intensive to determine if the reasoning was accurate or explainable beyond reasonable doubt; a definite, technical conclusion is often impossible (Bullock, 2020, p. 494). Truly analyzing a specific case might also infringe on other people's privacy (Magnusson Sjöberg, 2021), even though there are developments around transparency, openness, and intelligibility (Chakraborti et al., 2018). Outliers aside, diagnosing a specific case might thus be futile technologically and even more so in actual practice. This also means that, if at all, only well-resourced individuals—or paying customers—would be able to effectively challenge AI outcomes.

Complexity as a regulatory challenge

Complexity can be reduced in various, overlapping ways: first, the number of relevant entities can be reduced; for example, by removing unnecessary subsystems. Second, it is possible to reduce entity complexity (e.g., by streamlining code bases). Last, but not least, if the number of entities or systems cannot be reduced and their complexity is relatively set, standardization can be leveraged. For example, planes are highly complex but complexity is reduced by relying on a small number of plane types (De Florio, n.d., ch. 8). In the realm of AI, these approaches can be applied in some areas but are unlikely to be a perfect solution past development; AI systems must learn and change, being at least somewhat emergent in nature.

Furthermore, understanding and regulating AI requires insight into different aspects of a use case; this is possible but has not matured like in other sectors. Medical imaging analysis has proven to be a strong point of AI approaches, already leading to shifts in the

profession (Lewis et al., 2019). However, medical imaging involves a variety of aspects and actors, which influence who is being involved in the process and how, what tests are being done, how raw data are being analyzed, and what decisions are being made. For example, which processes lead to being selected for an (AI analyzed) scan in the first place? Patients themselves with their (medical) histories, doctors, technicians, AI systems, developers, hospital IT, but also medical coding and billing all play into how an image is read, analyzed, and acted upon. None of these actors have a full understanding of all relevant aspects.

The passing of time will also play a major role (Gigerenzer, 2018): the use case of a system might change during its development or thereafter, new data points might be added, or new systems integrated. This means that an initial design or assessment ceases to be functional or accurate. In fast moving spaces like AI, it is possible that by the time a (snapshot of) a technology or approach is well understood, we have already moved on to another. A system's inaccuracies might even lead to conclusions that then "confirm" the initial error. For example, if one out of two groups with similar rates of criminality were more likely to be flagged by autonomous systems due to an accidental, technical error, and thus stopped more often, police will logically find more criminals among that group; this is without even accounting for already existing racial bias (Williams, 2020).

Understanding AI, as is often requested in the literature (Ouchchy et al., 2020)—and perhaps a requirement for building systems that fulfill relevant rules and regulations—is not impossible, but it is time and resource intensive. Instead of understanding one algorithm, we would have to understand the training data, different software suites (some open source, some closed source), how training data were/are ingested, how data and algorithms are updated, what the AI system is supposed to do, how results are used, how results are presented, how results are controlled, how results are interpreted, what role humans play in the process, what input the AI system is making, etc. As Gigerenzer (2018, p. 307) observes, there is also a "bias Yanfeng"; that is, "the tendency to see systematic biases in behavior even when there is only unsystematic error or no verifiable error at all." According to Gigerenzer's research, there is a tendency to pursue errors in human thinking while ignoring machine errors. There is a similar, essentially inverted, problem that could impact on our understanding and the regulation of AI: random noise or small numbers of "biased-looking" outputs could be interpreted as systematic. Determining the difference between noise and signal will also be influenced by economic and political factors; key power brokers and players, as well as public discourse might co-determine what categories or cases of unfairness are being recognized, discussed, or investigated.

AI Regulation

Having considered a number of potential problems and groups of issues, this chapter takes a highly inclusive approach to regulation, defining it as "management of systems alongside set rules." This definition includes laws, private regulation (voluntary or contractually enforced), government regulations, professional standards, and more. To be effective, any form of regulation, including self-regulation, must include enforcement tools and consequences for parties that fail to comply. Without a way to enforce behavior, regulation cannot work.

Some regulatory objectives are rather simple conceptually, like avoiding system failure or accidents. For example, regulators required passenger aircrafts crossing the Atlantic to have more than two engines to increase redundancy in case of individual engine failure. Due to better technology (and agreed upon standards), many contemporary 2-jets are now cleared for such crossings (DeSantis, 2013). Even though airplanes and air traffic are complex systems, the safety objective and resulting requirements are relatively simple, revolving around the maximum distance/time away from a diversion airport.

When it comes to AI, objectives tend to be more complicated: transparency, explainability, fairness, or non-discrimination cannot be put into numbers as easily. What fairness means is also contested, with different, incompatible approaches and definitions being used (Verma & Rubin, 2018; Saxena et al., 2019). Furthermore, these objectives might indeed be different depending on system, use case, and situation. Thus, regulatory objectives pertaining to AI are complex in themselves. They require interpretation to become practical, actionable standards or guidelines, and must be adapted situationally. Furthermore, different jurisdictions may not arrive at the same conclusions, creating an assemblage of similar and dissimilar rules existing side-by-side. Having to address differing local, regional, or country-specific rules about AI, data use, or other relevant dependencies further adds to the complexity and brittleness of AI systems.

Different levels of regulation are likely to overlap in practice and they can be deployed alongside each other.[3] These levels of regulation do not speak to who is regulating; however, we tend to see higher-level issues covered by statutes, while more technical factors are in the realm of professional organizations and specialized (government) regulators. These levels of regulation could also be targeted differently; for example, by focusing on the creation, operation, or governance of AI systems.

1. Guiding operations and technology through specific, technical regulation, such as how a system component must be set up or maintained.
2. Changing the technology architecture or organizational approach, such as the creation of governance frameworks or specifying how certain systems are to be designed.
3. Providing high-level guidelines (values, objectives, etc.) to AI systems generally, be this through laws or by requiring organizations to create AI policies alongside given requirements.
4. Legislate around larger issue spaces, including AI in the process; this would likely be related to human rights and societal ills, capturing AI alongside other technologies.
5. To refrain from AI in certain circumstances. Such decisions might be made based on human rights concerns or because the risks outweigh potential gains.

Arguably, the scope of regulation needs to, at least in part, match the scope or realm of the problem it intends to address. Explicitly technical vulnerabilities or weaknesses are likely best served by technical guidelines, rather than national-level governance. Going up the "stack," would be more suitable to address how organizations use and govern AI architectures and systems overall.

However, turning a society-level concern into a set of engineering rules is difficult (Tampe, 2021). Implementation is necessarily local and undertaken by technical specialists who require high-level objectives to be translated into actionable and measurable standards.

The inherently political nature of AI regulation (Bartley, 2021) will also factor into any regulatory approach. Thus, fixing a societal ill is not possible with AI regulation. Even if a regulatory tool was perfectly successful in its niche, institutions, individuals, and other systems would not change.

Lessons from safety regulation?

Safety regulations in the airline and rail spaces have used the fact that while there are many individually complex entities to regulate, consistency across large groups of entities is given. Edge cases aside, plane families share many key objectives, features, and behaviors, allowing airworthiness to be decided based on type (De Florio, n.d.). In the aviation and, to a lesser extent, automotive and rail spaces, regulators also cooperate across borders in a variety of ways. For AI, this approach will likely become necessary but insufficient. Not only is AI considered a technology race and competition (Horowitz, 2018) but the whole idea of AI essentially requires all systems to be different from each other: they are meant to adapt, learn, and change. Thus, standardization could establish baselines regarding development and deployment but have less influence on system evolution.

The safety and AI spaces also present different problems: while doubtlessly hard, many objectives and issues in the safety space are "small world" problems. They are not easy to solve but the variables are known and consistent over time. In the AI space, "large world" issues are increasingly common: we (currently) know too little to design around a mostly complete set of risks, and instead must address uncertainty.

In contrast to a goal of "don't crash," aims like fairness or explainability are not only open to interpretation but require interpretation on a case-by-case basis—and that is difficult. To deal with this challenge, multi-layered governance structures will be necessary, where high-level objectives are created and then implemented organizationally by a variety of specialists, "trickling down" to the technical layers.

Nevertheless, decreasing uncertainty and risk has proven possible and effective. Even with the limitations imposed by the complexity of AI, progress is possible. However, this approach relies on experience and incidents to study, thus enabling learning over time. Aviation safety is high today partly because of decades of experience and incidents from which relevant regulators and engineers could learn. Cataloging incidents, as demonstrated by the Artificial Intelligence Incident Database[4] does help (McGregor, 2020) but time is of the essence in a fast moving space like AI.

Last, but not least, AI regulation faces an issue we see in other policy spaces: while incentives are not completely misaligned in all cases, there is a strong tendency for public interest to be misaligned with often powerful, well-resourced, and politically connected AI owners. Under the assumption that organizations and management want to pursue fairness or privacy, doing so will consume resources and potentially decrease the value of the AI system. At best, voluntarily pursuing those goals would likely not increase value, while performing the necessary implementation work would be costly. Thus, without regulatory pressure, pursuing goals like fairness would often be irrational for economically self-interested actors. Based on these limitations, this chapter proposes a multi-pronged approach to regulation, noting the importance of incentivization.

How to Regulate AI?

Regulation requires clarity

Regulations should ideally require appropriately defining AI and machine learning (ML) and categorizing AI systems alongside various relevant variables: for example, risk levels, use cases, autonomy, immediate impact, and complexity. One key element would be the consideration of how "intelligent," functional, or effective an AI or system would have to be to fall under (certain sections of) regulatory frameworks, requiring an operationalizable framework to define intelligence in the context of AI (Pfeifer & Scheier, 2001). Narrower AI likely requires less scrutiny, unless it is used in high-risk spaces. Complexity, the subject of this chapter, should also be a relevant measure. The more nodes (subsystems, external data sources, etc.) involved in a system, the higher its risk classification should become. While the creation of such tools is underway (Gasser & Almeida, 2017), the problem with this fundamental requirement is that we do not currently have definitions and measures that are agreed upon and sufficiently actionable for practitioners.

Regulation should be international and target organizations

On a high level, many factors play into outcomes created by AI systems. The more complex the regulated space or entity, the more likely it becomes that gray areas and disputes emerge; one could argue they are in themselves an aspect of emergence in complex systems (Goldstein, 2011). Thus, it is rather likely that one regulator would arrive at different definitions, indicators, and requirements than another. Privacy practitioners are facing this issue (Vasalou et al., 2015): different regimes raise compliance costs (Chander et al., 2021), workload, and uncertainty, usually being a boon to large, established players, and thus potentially stifling competition and innovation.

As with digital data flows and privacy laws, the global flow of data and information will complicate AI operations considerably. For example, how can digital rights, like the right to be forgotten, be implemented in AI systems where individual records could likely not be removed in full (Villaronga et al., 2018)? To answer some of these challenges, it is not unlikely that various AI systems as they exist today would require significant modification to be compliant with current rules, not even potential future regulatory regimes.

Therefore, legislators and regulators would benefit immensely from international cooperation in the AI space, which will likely be complicated by AI being seen as a competition and the logistical difficulty of designing international regulation.

Furthermore, regulation should target AI creators and users, establishing liability and requirements for them to implement internally, including a mandate for human-controlled exception handling procedures, and directives regarding transparency and human oversight over automated systems. The integration of AI systems into a wider set of controls and processes (Bullock, 2020, p. 503) could mitigate negative effects and control for error. Such measures are, at least in part, necessarily local and organizational. An organization will better understand their specific systems, niche, sector, and use case than

the legislative branch of government. Moreover, the social worker on the street, the diagnostician supporting a patient, the engineer physically at the plant, but also AI system administrators and programmers, usually have insights—as well as the capability for direct action—that AI systems, as well as distant bureaucrats or managers, are lacking. These professionals must retain agency and be given the tools to address challenges case-by-case, and to trigger change in AI systems to avoid future error. Implementing such management structures and controls that address fairness and other objectives of ethical or harm-reduced AI will be costly. While some actors would address these challenges willingly, others will not do so without external pressure.

The need for vertical regulation

Regulation usually creates negative externalities and unintended consequences, particularly under conditions of complexity: for example, diverging efforts away from important tasks, compliance culture with little positive effect, or regulation gaming. All these potential dangers are fanned by complexity: the more variance there is in approaches and systems, and the more work required to deal with them, the harder it becomes to create rules that are befitting for all the systems, as well as efficiently and effectively implementable.

In the security space, the tools to achieve the goals of confidentiality, integrity, and availability are multi-layered. Legal requirements are increasingly common but insurers and other relevant parties have also stipulated requirements. Organizations must, in many cases, adopt a governance approach, defining policies at the top that are in line with both relevant laws and their specific case, and then through a variety of steps, apply them to technology on the ground (Peltier, 2016). Essentially, to capture diverging AI systems, organizations must go through a similar process of defining their AI policies, procedures, architectures, guidelines, and more, in accordance with relevant laws and regulations. While resource-intensive, this appears to be the only feasible approach that can address the complexity challenges. For example, it is only logical to argue that a fully autonomous system should face more scrutiny than an advisory one. However, to do this accurately, professionals need standards to measure and classify systems that are applicable in practice and appropriate for their environment.[5] Similarly, standards on measuring complexity, fairness, and other key variables will be needed, given that there is pressure from government or other actors.

However, an immediate requirement is appropriate reporting and transparency (Waltl & Vogl, 2018). Not only is detailed, accurate, and understandable documentation important for individual observers and regulators, but the process of creating such documentation is useful for system owners and architects.

The need for AI specialists

Due to the intricacy of AI systems in their technical, social, and organizational contexts, any approach discussed in this chapter requires considerable expertise by regulators, system architects, and other professionals. Indeed, commercial and civil society expert groups that look into AI systems have emerged—but with more, increasingly complex AI being deployed, the need for experts will grow.

As with cyber security professionals, it is likely that demand is outpacing the supply of AI specialists in general (Oomen, 2021; Zwetsloot, 2019), and in particular those that have a deep grasp of the overall technology stack. Mandatory training for AI system operators and users might also be required (e.g., Mazurowski, 2019; Abeliuk et al., 2020). The more widespread such systems become, the more people would have to interpret outputs, maintain systems, diagnose errors, and act in the capacity of moderator.

This means that as one of the first steps, lawmakers and other interested parties should foster talent and skill pipelines that focus on technical skill but also include mandatory training in AI ethics, management processes, and focus on the overall understanding and analysis of AI in all their facets and impacts.

Why AI regulation "needs teeth"

In the context of larger social ills, like inequality, racism, and, importantly, a lack of (corporate) accountability, the political economy of AI matters. Crucially, the users of AI will have to be made responsible for what their systems "create" in terms of outcomes. Regulation only works when there are consequences to breaking the rules or causing harm. Obviously, assigning blame and liability will not always be easy but HIPAA (and other regulations) solve such an issue by creating liability for multiple parties like "business associates,"[6] thus incentivizing different parties to engage in due care and due diligence (Department of Health and Human Services, 2013). As in other spaces, legal liability and responsibility will elicit the creation of tools to address liability and regulatory requirements. In complex system computing, approaches to make complex systems function more reliably were required, and thus developed.

The creation of rules will likely lead to greater monopolization with the largest players being able to best absorb the costs (Roberts, 2018). Those players will also use very complicated systems that badly resourced regulators would struggle to readily understand, similar to other spaces where the regulated often have more insight, data, information, and crucially, resources than those meant to control them (Bloom et al., 2014). That said, if regulators do have the ability to punish, and choose to apply this incentive, their work will contribute to less risky behavior and harm reduction. As in IT Security, even an organization that truly tries to address all threats and concerns will see AI failures. Thus, similarly, governance, risk reduction, and management procedures are likely to become requirements. Establishing proof of due care and due diligence will reduce liability if, or rather, when things go wrong.

As with the General Data Protection Regulation (European Commission, 2016), the EU acted early regarding AI regulation (European Commission, 2021), while clearly lagging behind the United States and China in development (Oomen, 2021). Regulation can serve as a way to control foreign influence and power, and this is arguably the case here. On the other hand, under-regulation and laissez-faire approaches may support quicker, more economical progress at the cost of greater social harm. Arguably, some (future) uses of AI might be costing us more as a society than they provide. Indeed, such considerations might be driving the EU approach, at least to some extent (Brattberg et al., 2020). As Gigerenzer (2018) observes, humans are better as decision makers than they are given credit for. Thus,

AI might not always bring significant (enough) benefits to warrant their resource consumption and potential for harm.

In sum, regulation needs to be balanced: it is only rational for governments to act, especially if many uses of AI in the relevant territory are undertaken by unaccountable (foreign) players and/or if their use cases are considered unethical or opposed to relevant human or citizens' rights. However, over-regulation is likely to impact negatively on progress and might worsen the harms we are currently observing. Most of all, there has to be the political will (internationally) to regulate and enforce relevant rules. Governments and regulators must introduce strong incentives, including legal liability and impactful fines for companies and their relevant executives when their AI systems produce harm or are not created or managed with due care and due diligence.

Discussion and Conclusion

This chapter has discussed regulation of AI systems from a perspective that focuses less on what to regulate but rather on how this space complicates and enables regulation and governance. The inherent complexity of computer technology and data flows, culminating in high-complexity AI systems, poses a major challenge to regulators; as with many other technologies, grasping dependencies and how systems actually function is non-trivial. Achieving sufficient grasp of whole systems involves technical but also organizational, niche, and systemic understanding.

This does not mean that meaningful regulation of AI is impossible or unnecessary, but rather that it requires well-crafted, constantly updated rules. Indeed, there should be a measure of system complexity that identifies at-risk systems. To achieve this, a multi-layered system of laws (slow and high level), standards, and best practices will be required, similar to what we find in other sectors like information security. Unfortunately, such multi-layered solutions have drawbacks, such as overhead, compliance culture, and conflicting objectives and interpretations.

Many of the harms discussed in the context of AI are linked to larger societal ills. Some noteworthy examples aside, AI usually did not create the problems we identify but makes them worse or more apparent. Even in cases where AI does create new problems, like in the case of facial recognition, these problems are embedded in more macro-level social concerns like an ever-expanding and unaccountable digital surveillance machine. This is not to say that regulating AI-specific issues is redundant, but rather that some questions are likely better regulated on a more fundamental level with additional AI-specific rules being added on.

Powerful and well-resourced actors will likely continue using the most influential and powerful AI systems, while individuals will likely be on the receiving end. There is a clear and worrisome risk that individuals' data and their lives will be increasingly determined by largely unaccountable AI systems run by mostly unaccountable businesses.

This lack of accountability, liability, and (simple) recourse for those wronged should be the key objective for regulators and political institutions, while giving organizations some leeway in terms of how to design their systems and governance structures.

Acknowledgments

I would like to thank Baobao Zhang, Toby Shevlane, Gretchen Greene, and Tiffany Li who kindly commented on my draft. I would also like to thank the Members of the Yale Law School Information Society Project for their comments on an early concept of this paper.

Notes

1. For a more in-depth discussions of levels of complexity, see Weissinger (2020).
2. Where different studies come to different conclusions about who is being favored, depending on situation and key variables (Carlsson and Eriksson, 2019; Grant & Mizzi, 2014; Isaac et al., 2009).
3. As highlighted by the EU AI regulation proposal (European Commission, 2021).
4. See https://incidentdatabase.ai/.
5. As Abeliuk et al. (2020) and Schmidt and Biessmann (2020) find, how AI are integrated into workflows remains multifaceted.
6. See https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/business-associates/index.html.

References

Abeliuk, Andrés, Benjamin, Daniel M., Morstatter, Fred, & Galstyan, Aram. (2020). Quantifying machine influence over human forecasters. *Scientific Reports 10* (1), 15940. http://www.nature.com/articles/s41598-020-72690-4.

Agrawal, Tushar. (2014). Gender and caste-based wage discrimination in India: some recent evidence. *Journal for Labour Market Research 47* (4), 329–340. http://link.springer.com/10.1007/s12651-013-0152-z.

Alexander, Alan, Jiang, Adam, Ferreira, Cara, & Zurkiya, Delphine. (2020). An intelligent future for medical imaging: A market outlook on artificial intelligence for medical imaging. *Journal of the American College of Radiology 17* (1), 165–170. https://linkinghub.elsevier.com/ retrieve/pii/S1546144019308634.

Anton, Eduard, Behne, Alina, & Teuteberg, Frank. (2020). The humans behind artificial intelligence—An operationalisation of AI competencies. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, 19. Marrakech.

Banerjee, Debarag Narayan, & Chanda, Sasanka Sekhar. (2020). AI failures: A review of underlying issues. *arXiv:2008.04073 [cs]*. http://arxiv.org/abs/2008.04073.

Barocas, Solon, & Selbst, Andrew D. (2016). *Big data's disparate impact*. SSRN Scholarly Paper ID 2477899. Social Science Research Network. https://papers.ssrn.com/abstract=2477899.

Bartley, Tim. (2021). Power and the practice of transnational private regulation. *New Political Economy*, 1–15. https://doi.org/10.1080/13563467.2021.1881471.

Beck, Ulrich. 1986. *Risikogesellschaft. Auf dem weg in eine andere moderne*. Suhrkamp.

Beck, Ulrich, & Beck-Gernsheim, Elisabeth. (1994). *Riskante freiheiten*. Suhrkamp.

Beyer, Betsy, Jones, Chris, Petoff, Jennifer, & Murphy, Niall Richard. (2016). *Site reliability engineering: How Google runs production systems*. O'Reilly Media.

Bloom, Gerald, Henson, Spencer, and Peters, David H. (2014). Innovation in regulation of rapidly changing health markets. *Globalization and Health 10* (1), 53. https://doi.org/10.1186/1744-8603-10-53.

Brattberg, Erik, Rugova, Venesa, & Csernatoni, Raluca. (2020). *Europe and AI: Leading, lagging behind, or carving its own way?* Carnegie Endowment for International Peace.

Bullock, Justin. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, Preprint, 1–15.

Carlsson, Magnus, & Eriksson, Stefan. (2019). In-group gender bias in hiring: Real-world evidence. *Economics Letters 185* , 108686. https://www.sciencedirect. com/science/article/pii/S0165176519303428.

Chakraborti, Tathagata, Kulkarni, Anagha, Sreedharan, Sarath, Smith, David E., & Kambhampati, Subbarao. (2019). *Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior*. In *Proceedings of the International Conference on Automated Planning and Scheduling* (vol. 29, pp. 86–96).

Chander, Anupam, Abraham, Meaza, Chandy, Sandeep, Fang, Yuan, Park, Dayoung, & Yu, Isabel. (2021). Achieving privacy: Costs of compliance and enforcement of data protection regulation. In *Policy Research Working Paper 9594. World Bank's World Development Report 2021* (pp. 1–42). World Bank. https://www.ssrn.com/abstract=3827228.

Chen, Benjamin Minhao, Stremitzer, Alexander, & Tobia, Kevin. (2021). *Having your day in robot court*. SSRN Scholarly Paper ID 3841534. Social Science Research Network. https://papers.ssrn.com/abstract=3841534.

De Florio, F. (n.d.). *Airworthiness: An introduction to aircraft certification and operations*. Elsevier Science.

Department of Health and Human Services. (2013). Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the health information technology for economic and clinical health act and the genetic information nondiscrimination act; other modifications to the HIPAA rules. *Federal Register 78* (17), 5566–5702. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/combined-regulation-text/omnibus-hipaa-rulemaking/ index.html.

DeSantis, J. Angelo. (2013). Engines turn or passengers swim: A case study of how ETOPS improved safety and economics in aviation. *Journal of Air Law and Commerce 77* (4), 67.

Dressel, Julia, & Farid, Hany. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4* (1), 1–5. https://advances.sciencemag.org/content/4/1/eaao5580.

Durkheim, Émile. (1984). *The division of labour in society*. New edition. Palgrave Macmillan.

European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union, L 119*. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Commission. (2021). COM/2021/206 final: Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *Official Journal of the European Union*, COM(2021) 206 final 2021/0106(COD). https://eur-lex.europa.eu/legal content/EN/ALL/?uri=CELEX:52021PC0206.

Ferrer, Xavier, van Nuenen, Tom, Such, Jose M., Coté, Mark, & Criado, Natalia. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine* 40 (2), 72–80.

Fine Licht, Karl de, & de Fine Licht, Jenny. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & SOCIETY 35* (4), 917–926. http://link.springer.com/10.1007/s00146-020-00960-w.

Foucault, M. (2012). *Discipline and punish: The birth of the prison*. Vintage. Knopf Doubleday Publishing Group.

Gardenier, John, & Resnik, David. (2002). The misuse of statistics: Concepts, tools, and a research agenda. *Accountability in Research 9* (2), 65–74. https://doi.org/10.1080/08989620212968.

Gasser, Urs, & Almeida, Virgilio A. F. (2017). A layered model for AI governance. *IEEE Internet Computing 21* (6), 58–62. http://ieeexplore.ieee.org/document/8114684/.

Gigerenzer, Gerd. (2018). The bias bias in behavioral economics. Review of *Behavioral Economics 5* , 303–326.

Gigerenzer, Gerd, & Gaissmaier, Wolfgang. (2011). Heuristic decision making. *Annual Review of Psychology 62* (1), 451–482. http://www.annualreviews.org/doi/10.1146/annurev-psych-120709-145346.

Gillespie, Tarleton. (2020). Content moderation, AI, and the question of scale. *Big Data & Society 7* (2), 205395172094323. http://journals.sagepub.com/doi/10.1177/2053951720943234.

Goldstein, Jeffrey. (2011). Emergence in complex systems. In Peter Allen, Steve Maguire, & Bill McKelvey (Eds.), *The SAGE handbook of complexity and management* (pp. 65–78). SAGE.

Grant, Sharon, & Mizzi, Toby. (2014). Body weight bias in hiring decisions: Identifying explanatory mechanisms. *Social Behavior and Personality: An International Journal 42* (3), 353–370.

Hill, Kashmir. (2020a). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*, December 29. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

Hill, Kashmir. (2020b). Wrongfully accused by an algorithm. *The New York Times*, June 24. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hirsch, Paul, Fiss, Peer C., & Hoel-Green, Amanda. (2009). A Durkheimian approach to globalization. In P. Adler (Ed.), *The Oxford Handbook of Sociology and Organization Studies: Classical Foundations* (pp. 223–245). Oxford University Press. http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199535231.001.0001/oxfordhb-9780199535231-e-010.

Horowitz, Michael C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review 1* (3), 22.

Isaac, Carol, Lee, Barbara, & Carnes, Molly. (2009). Interventions that affect gender bias in hiring: A systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges 84* (10), 1440–1446. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4554714/.

Kaur, Paramjit, et al. (2020). Facial-recognition algorithms: A literature review. *Medicine, Science and the Law 60* (2), 131–139. http://journals.sagepub.com/doi/10.1177/0025802419893168.

Lahtinen, Hannu, Sirniö, Outi, & Martikainen, Pekka. (2020). Social class and the risk of unemployment: Trends, gender differences and the contribution of education. *Acta Sociologica 63* (3), 303–321. http://journals.sagepub.com/doi/10.1177/0001699318817594.

Larsson, Stefan, & Heintz, Fredrik. (2020). Transparency in artificial intelligence. *Internet Policy Review 9* (2), 1–16. https://policyreview.info/node/1469.

Lee, Erika. (2021). *America for Americans: A history of xenophobia in the United States*. Basic Books.

Lewis, Sarah J., Gandomkar, Ziba, & Brennan, Patrick C. (2019). Artificial intelligence in medical imaging practice: Looking to the future. *Journal of Medical Radiation Sciences 66* (4), 292–295. https://onlinelibrary.wiley.com/doi/abs/10.1002/jmrs.369.

Llansó, Emma, et al. (2020). Artificial intelligence. *Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, 32.

Magnusson Sjöberg, Cecilia. (2021). Legal AI from a privacy point of view: Data protection and transparency in focus. In Sonya Petersson (Ed.), *Digital human sciences: New objects—new approaches* (pp. 181–202). Stockholm University Press. https://www.stockholmuniversitypress.se/site/chapters/e/10.16993/bbk.h/.

Manheim, Karl, & Kaplan, Lyric. (2019). Artificial intelligence: Risks to privacy and democracy. *Yale Journal of Law and Technology 21* , 83.

Mazurowski, Maciej A. (2019). Artificial intelligence may cause a significant disruption to the radiology workforce. *Journal of the American College of Radiology 16* (8), 1077–1082. https://linkinghub.elsevier.com/retrieve/pii/S154614401930064X.

McGregor, Sean. (2020). Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *arXiv:2011.08512 [cs]*. http://arxiv.org/abs/2011.08512.

Michel, Arthur Holland. (2021). *Known unknowns: Data issues and military autonomous systems*. United Nations Institute for Disarmament Research.

Narayanan, Arvind. (2019). How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics, Massachusetts Institute of Technology*. https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

Oomen, Thomas L. (2021, February 22). Why the EU lacks behind China in AI development. *Article 2021-8, Asia Centre*, 17.

Ouchchy, Leila, Coin, Allen, & Dubljević, Veljko. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY 35* (4), 927–936. http://link.springer.com/10.1007/s00146-020-00965-5.

Pasquale, Frank, & Malgieri, Gianclaudio. (2021, August 2). If you don't trust A.I. yet, you're not wrong. *New York Times*, Section A, Page 19. https://www.nytimes.com/2021/07/30/opinion/artificial-intelligence-european-union.html.

Pawson, Ray, Wong, Geoff, & Owen, Lesley. (2011). Known knowns, known unknowns, unknown unknowns: The predicament of evidence-based policy. *American Journal of Evaluation 32* (4), 518–546. http://journals.sagepub.com/doi/10.1177/1098214011403831.

Peltier, T. R. (2016). *Information security policies, procedures, and standards: Guidelines for effective information security management*. CRC Press.

Pesapane, Filippo, Codari, Marina, & Sardanelli, Francesco. (2018). Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental 2* (1), 35. https://eurradiolexp.springeropen.com/articles/10.1186/s41747-018-0061-6.

Pfeifer, R., & C. Scheier. (2001). *Understanding intelligence*. MIT Press. https://books.google.com/books?id=iIv6-rxTCkoC.

Povolny, Steve, & Trivedi, Shivangee. (2020). Model hacking ADAS to pave safer roads for autonomous vehicles. *McAfee Blogs*. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/.

Rescher, Nicholas. (2020). *Complexity: A philosophical overview*. 1st ed. Routledge. https://www.taylorfrancis.com/books/9781000659474.

Roberts, Sarah T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*. https://www.firstmonday.org/ojs/index.php/fm/article/view/8283.

Röösli, Eliane, Rice, Brian, and Hernandez-Boussard, Tina. (2021). Bias at warp speed: How AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association 28* (1), 190–192. https://academic.oup.com/jamia/article/28/1/190/5893483.

Ropek, Lucas. (2021). Report: Cops have been pressuring ShotSpotter to alter evidence. *Bloomberg*. https://gizmodo.com/report-cops-have-been-pressuring-shotspotter-to-alter-1847362168.

Roten, Fabienne Crettaz von. (2006). Do we need a public understanding of statistics? *Public Understanding of Science 15* (2), 243–249. https://doi.org/10.1177/0963662506061883.

Saleiro, Pedro, et al. (2019). "Aequitas: A bias and fairness audit toolkit". *arXiv:1811.05577 [cs]*. http://arxiv.org/abs/1811.05577.

Sánchez-Monedero, Javier, Dencik, Lina, & Edwards, Lilian. (2020). What does it mean to "solve" the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York: Association for Computing Machinery, 458–468. https://doi.org/10.1145/3351095.3372849.

Saxena, Nripsuta Ani, et al. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 99–106. https://dl.acm.org/doi/10.1145/3306618.3314248.

Schelling, Thomas. (2006). *Micro motives and macro behavior*. 1st edition. Norton.

Schmidt, Philipp, and Biessmann, Felix. (2020). Calibrating human–AI collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In Andreas Holzinger et al. (Ed.), *Machine learning and knowledge extraction*, *vol. 12279* (pp. 431–449). Springer International Publishing, http://link.springer.com/10.1007/978-3-030-57321-8_24.

Selinger, Evan, and Hartzog, Woodrow. (2020). *The inconsentability of facial surveillance*. SSRN Scholarly Paper ID 3557508. Social Science Research Network. https://papers.ssrn.com/abstract=3557508.

Skitka, Linda J., Mosier, Kathleen L., and Burdick, Mark. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies 51* (5), 991–1006. https://www.sciencedirect.com/science/article/pii/S1071581999902525.

Soper, Spencer. (2021). Fired by bot at Amazon: "It's you against the machine." *Bloomberg*. https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out.

SteelFisher, Gillian K., et al. (2019). Gender discrimination in the United States: Experiences of women. Health Services *Research 54* (S2), 1442–1453. https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.13217.

Tampe, Maja. (2021). Turning rules into practices: An inside-out approach to understanding the implementation of sustainability standards. *Ecological Economics 184* , 106947. https://linkinghub.elsevier.com/retrieve/pii/S0921800921000057.

Tolba, A. S., El-Baz, A. H., & El-Harby, A. A. (2008). Face recognition: A literature review. *International Scholarly and Scientific Research & Innovation 2* (7), 16.

Vasalou, Asimina, Joinson, Adam, & David Houghton. (2015). Privacy as a fuzzy concept: A new conceptualization of privacy for practitioners. *Journal of the Association for Information Science and Technology 66* (5), 918–929. http://doi.wiley.com/10.1002/asi.23220.

Verma, Sahil, & Rubin, Julia. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering. ACM, 1–7. https://dl.acm.org/doi/10.1145/3194770.3194776.

Villaronga, Eduard Fosch, Kieseberg, Peter, & Li, Tiffany. (2018). Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review 34* (2), 304–313. https://linkinghub.elsevier.com/retrieve/pii/S0267364917302091.

Wachter, Sandra, Mittelstadt, Brent, & Russell, Chris. (2020). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *SSRN Electronic Journal arXiv: 2005.05906*. http://arxiv.org/abs/2005.05906.

Waltl, Bernhard, & Vogl, Roland. (2018). Increasing transparency in algorithmic-decision-making with explainable AI. *Datenschutz und Datensicherheit—DuD 42* (10), 613–617. http://link.springer.com/10.1007/s11623-018-1011-4.

Weissinger, Laurin B. (2020). The challenge of networked complexity to NATO's digital security. In *Cyber Threats and NATO 2030: Horizon Scanning and Analysis*, 17. https://ccdcoe.org/uploads/2020/12/13-The-Challenge-of-Networked-Complexity-to-NATOs-Digital-Security_ebook.pdf.

Williams, Damien Patrick. (2020). Fitting the description: Historical and sociotechnical elements of facial recognition and anti-black surveillance. *Journal of Responsible Innovation 7* (sup1), 74–83. https://www.tandfonline.com/doi/full/10.1080/23299460.2020.1831365.

Woodward, Mark, Peters, Sanne A. E., & Harris, Katie. (2021). Social deprivation as a risk factor for COVID-19 mortality among women and men in the UK Biobank: Nature of risk and context suggests that social interventions are essential to mitigate the effects of future pandemics. *Journal of Epidemiology and Community Health 75* , 1050–1055. https://jech.bmj.com/lookup/doi/10.1136/jech-2020-215810.

Young, Matthew M., et al. (2019). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4* (3), 244–258.

Zhang, Yunfeng, Bellamy, Rachel, & Varshney, Kush. (2020). Joint optimization of AI fairness and utility: A human-centered approach. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, ACM, 400–406. https://dl.acm.org/doi/10.1145/3375627.3375862.

Zuboff, Shoshana. (2019). Surveillance capitalism and the challenge of collective action. *New Labor Forum 28* (1), 10–29. http://journals.sagepub.com/doi/10.1177/1095796018819461.

Zwetsloot, Remco. (2019). *Strengthening the U.S. AI workforce*. Center for Security and Emerging Technology. https://cset.georgetown.edu/research/strengthening-the-u-s-ai-workforce/.

SECTION VII

ECONOMIC DIMENSIONS OF AI GOVERNANCE

ANTON KORINEK

CHAPTER 33

TECHNOLOGICAL UNEMPLOYMENT

DANIEL SUSSKIND

Introduction

In the early 1980s, Wassily Leontief, a Nobel-Prize-winning economist, made one of the most provocative claims in economic thought. He feared that what cars and tractors had done to horses at the turn of the 19th century, computers and robots would eventually do to human beings as well—drive more and more of us out of paid work (Leontief, 1979, 1983a, 1983b). And today, the world is captured again by Leontief's fear: in the U.S., 30 percent of workers now believe their jobs are likely to be replaced by robots and computers in their lifetime; in the UK, the same proportion think it could happen in the next 20 years (Susskind, 2020a).

At the time Leontief was writing, his view was not mainstream in the economics profession. And in the decades that followed, it has proven to be very difficult to engage with his concerns in the core models that explore the impact of technology on the labor market.

These models, designed to explain certain empirical puzzles that were emerging in labor markets at the time they were conceived, tended to support a far more optimistic view of the impact of technological change on the labor market. There was little room for Leontief's fears. But in the more recent economic literature, there has been a distinct change. New empirical features of the labor market, and striking developments in the capabilities of the latest technologies, have resulted in the development of a new set of core models. These now do support the possibility of a more pessimistic account of the impact of new technologies.

In this chapter, I explore this history of economic thought that focuses the impact of technological progress on the labor market. I look at the core models that have been developed, describing the succession of empirical puzzles that they were trying to explain and discussing their respective strengths and weaknesses. My aim is to chart the general change of heart that I detect in the economics literature over the past 80 years or so—from a blinkered optimism about the impact of technological change on the labor market, to a creeping pessimism. I conclude that these intellectual developments mean we now can and ought to

take the spirit—though not always the substance—of Leontief's fears more seriously than we might have done when he first wrote them down four decades ago.

Technology and The Labor Market: A Brief History

The neoclassical production function

An important starting point in the history of economic thought that explores the impact of technological progress on the labor market is the traditional "neoclassical production function." Here, the economy is captured by a single function that describes how different factors combine to produce output. And in this setting, new technologies can do three things: they can make labor more productive, so that it is as if there is more labor in the economy; this is "labor-augmenting" technological change (Harrod, 1948). They can make capital more productive, so it is as if there were more capital; this is "capital-augmenting" technological change (Solow, 1956). Finally, they can do both (Hicks, 1932).

To economists, this approach is simple and familiar. But it is important to reflect on how it might have constrained subsequent economic thinking. And very broadly, there are two main concerns with this basic set-up. First, it focuses our attention almost exclusively on how technology might "help" a particular factor—technological progress only appears as an "augmenting" or "complementing" force, and there is no way to capture the idea that new technologies might displace workers from a growing range of tasks (Acemoglu & Restrepo, 2018a make a similar complaint). In turn, this "aggregate" approach, which captures the economy in a single mathematical function alone, is opaque on how new technologies affect different factors—the impression is that new technologies simply increase the productivity of the relevant factor at *everything* that it does, leading to more output as a result. Both these features have reappeared in later literature in different ways, as we shall see, and unhelpfully cloud of view of how new technologies might harm workers—now, or in the future.

Skill-biased technical change

Until the turn of the 20th century, the "skill-biased technical change" (SBTC) thesis was the main way that economists thought about the impact of technological progress on the labor market. One simple exposition of the SBTC thesis was captured in what Acemoglu and Autor (2011, 2012) call the "canonical model": this uses the neoclassical production function from before, with two different types of labor, low-skilled and high-skilled, which combine to produce output in a carefully specified way (through a so-called "constant elasticity of substitution" production function, meaning that a percentage change in the relative wage of the different types of workers always causes a constant percentage change in the relative use of those workers).

The SBTC thesis was designed to explain an empirical puzzle—that during the post-war period the supply of educated workers in many labor markets increased but, at the

same time, the "price" (i.e., the wage) of those workers rose rather than fell relative to those without an education (see, for instance, Acemoglu, 2002). The SBTC thesis explained this by arguing that technological change was "skill-biased"—new technologies (like the digital electronic computer) required skilled people to put them to effective use, and this increased the demand for skilled workers to such an extent that, although their supply had increased, so did their relative price, the so-called "skill premium" (i.e., the wage of an average college graduate relative to an average high school graduate).

The theoretical and empirical literature on the SBTC thesis is now very large (see, for instance, Katz & Murphy, 1992; Bound & Johnson, 1992; Goldin & Katz, 1998; Bekman et al., 1998; Autor et al., 1998; Card & Lemieux, 2001; Acemoglu, 2002 noted before; and Goldin & Katz, 2008). However, although the SBTC thesis is useful in explaining that distinctive rising skill premium, two problems are associated with the wider use of the canonical model in thinking about the impact of technology on work. Both issues are rooted in its reliance on the neoclassical production function. The first is that, in the canonical model, there is no way for technological progress to make either type of worker worse-off in absolute terms—the relative wage of a worker can fall, but not their absolute wage (i.e., technological change is necessarily a "q-complement" to both types of labor; see Susskind, 2020c for a broader discussion of this point). From an empirical point of view, this is a problem because certain workers have already been made worse-off (see, for instance, Autor, 2019 on the declining wages of non-college workers). The second problem is that the canonical model is semantically deprived—there is no formal way to understand how or why it is that technological change makes high-skilled workers better at their work. As Autor et al. (2003) put it, "[i]t fails to answer the question of what it is that computers do—or what it is that people do with computers—that causes educated workers to be relatively more in demand."

Routine-task-replacing technical change

Autor et al. (2003) was originally intended to provide a deeper explanation for the SBTC thesis. But in doing so, it also introduced two innovations into the literature exploring the impact of technology on the labor market: the so-called "task-based approach," and an influential theory of the capabilities of systems and machines. Together, these have become known as the "Autor-Levy-Murnane" hypothesis ("the ALM hypothesis"). This line of thinking has been developed further in more recent research (see, for instance, Levy & Murnane, 2004; Autor, 2013, 2015a, 2015c).

The first innovation in the ALM hypothesis was the distinction between a "job" and the different "tasks" that make it up.[1] As noted, one of the problems with neoclassical production functions is that they provide no explanation for why or how technology affects different factors—factors combine to produce output, new technologies allow factors to produce more output, and that is that. We saw this lack of explanation with the canonical model: new technologies allow high-skilled workers to produce more output, but there is no reasoning that details why that is the case or how this process works. By introducing the concept of "tasks," the ALM hypothesis was able to provide this missing account. The neoclassical production function was broken down into two composite production functions: one that captured how different tasks combine to produce goods (i.e., a "task-based production for goods") and another that describes how different factors with different

capabilities combine to perform those tasks (i.e., a "factor-based production function for tasks"). When these two functions were combined, one would be able to recover the neoclassical production function once again. But this breakdown allowed for a deeper explanation of how technology affected a particular factor: new technologies helped factors by making them more productive at certain tasks but might harm them by displacing them from others (Susskind, 2016).

To see this deeper explanation in practice, think of the personal computer (PC). In the late 1950s and early 1960s, businesses started to make use of mainframe computers. Soon after that, adoption of the PC began: as late as 1980, the U.S. had fewer than one PC per 100 people, but by the turn of the century that figure had risen to more than 60. And these machines not only became more widespread but more powerful as well: on one estimate, for instance, computational power from 1950 to 2000 increased roughly by a factor of 10 billion (Susskind, 2020a). How does this technological progress relate to the changes that were taking place in the labor market at the time? The task-based approach provides a more detailed explanation: these powerful new machines led to an increase in demand for the particular types of *tasks* performed by high-skilled workers who were able to effectively operate these new technologies. This sort of explanation for SBTC, simple and intuitive it may be, could not be provided with a traditional neoclassical production function.

Building on this distinction between "jobs" and "tasks," the second innovation was a further distinction between "routine" and "non-routine" tasks. The argument was that while systems and machines could readily perform "routine" tasks, they would struggle to perform "non-routine" ones. The motivation for this claim relies on a particular conception of how systems and machines operate—that to perform a task, they must follow an explicit set of instructions or rules articulated by a human being. This very closely reflects the "expert systems" approach to AI, which was particularly influential in the 1980s (Winston, 1977; Hayes-Roth et al., 1983; and Russell et al., 1995 provide an overview of this technique). And, according to economists, a task was "routine" if a human being could articulate how they performed it—following Polanyi (1966), if perform the task relied on "explicit" rather than "tacit" knowledge. If this was the case, then it meant it was straightforward to write a set of instructions, based on that human explanation, for a machine to follow—and so these tasks could be automated. Otherwise, the tasks were "non-routine," the instructions or rules were inexpressible, and so they could not be automated. This distinction between "routine" and "non-routine" tasks is now commonplace (see Susskind, 2020a, 2020b, 2020c for a fuller history).

The ALM hypothesis was powerful because it could explain a different empirical puzzle that emerged in labor markets at the turn of the 20th century. Here, technological progress appeared to harm middling-skilled workers whose pay and share of jobs (as a proportion of total employment) declined in many countries. This phenomenon was known as the "polarization" or the "hollowing-out" of the labor market (see Goos & Manning 2007; Autor et al., 2006, 2008; Goos et al., 2009, 2014). And the most compelling explanation was provided by the ALM hypothesis—low- and high-skilled jobs were heavily composed of "non-routine" tasks which could not readily be automated but middling-skilled jobs were made up of "routine" tasks that were easier to automate. This reasoning could explain the hourglass figure that was emerging in many labor markets.

In time, though, a problem emerged with the ALM hypothesis: many of the tasks that it identified as being "non-routine," and out of reach of systems and machines, could

increasingly be automated. For instance, the tasks of making a medical diagnosis, driving a car, and identifying a bird at a fleeting glimpse were all thought to be out of reach—and yet there are now many systems that can make medical diagnoses, almost all major car manufacturers have driverless car programs, and there is even an app that can identify a bird at a quick glance (see Susskind & Susskind, 2015; Susskind, 2019, 2020a, 2020b, 2020c). Of course, it is reasonable to question whether these tasks have been fully automated: there are certain illnesses these diagnostic systems cannot identify; particular birds that the app cannot recognize; and no driverless car that can yet function without some human attention. But it is important to note the underlying trend: that many "non-routine" tasks can be automated to an extent that was inconceivable a few decades ago (Susskind, 2020a).

At fault was the underlying conception of machine capabilities on which the ALM hypothesis relied. As noted, this reflected how systems and machines worked in the "first wave" of AI that took place in the 1980s—the so-called "expert systems" approach—but this did not reflect the different approaches that were emerging at the turn of the century. Newer technologies, using advances in processing-power, data storage capabilities, and algorithm design, were no longer reliant on an explicit top-down articulation of instructions from human beings; instead, they were learning from the bottom-up (Susskind & Susskind, 2015; Susskind, 2020a). Take the system developed at Stanford that, from a photo, can diagnose whether a freckle is cancerous as accurately as 21 leading dermatologists (Esteva et al., 2017). How does it function? Not by following an explicit set of rules articulated by a human doctor, but by identifying patterns in a database of 129,450 past clinical cases. Of course, some of the rules it uncovers in that process may be those that a human doctor relies upon but struggles to articulate. But not necessarily so: the system may discover entirely new rules, too. As a result, in this way an increasing range of "non-routine" activity is within reach of these systems and machines, which operate in a very different way to human beings.[2]

Newer approaches

More recent approaches in the literature build on the combined shortcomings of the SBTC and the ALM hypothesis. Often, these continue to adopt the task-based approach, helping to resolve some of the analytical opacities that come with using neoclassical production functions. In turn, they are far more agnostic about the capabilities of machines, avoiding the risks of underestimating them as other researchers have done in the past. A good example of this "task-based, capabilities-agnostic" approach is Acemoglu and Restrepo (2018b).

This latter feature, that these models are often "capabilities-agnostic," is particularly important when thinking about the future of work. In these newer approaches, the explicit articulations of machine capabilities and functionality that characterized the earlier literature—the SBTC assumption that technology helps skilled-workers, or the ALM hypothesis that "non-routine" tasks cannot readily be automated—are gone. Instead, there is an implicit recognition that the boundaries to machine capabilities are uncertain and changing. In that spirit, many newer models instead use an endogenously determined cut-off in task space that marks the boundary between activities that can and cannot be automated (for instance, in Acemoglu & Restrepo, 2018b; Aghion et al., 2019; Moll et al., 2021). There is a reliance on the data, rather than theory, to identify where this moveable

boundary might lie—for instance, the AI Occupational Impact Measure, the Suitability for Machine Learning Index, and the AI Exposure Score (Acemoglu et al., 2021). And as technological progress unfolds, the cut-off in task space shifts—there is a process of "task encroachment" at work, where machines gradually, but relentlessly, take on more tasks (see Susskind, 2020a, 2020c).

Yet despite this agnosticism, there remains an important sense in which these new approaches still assume that labor is "more capable" than machines. It is true that the task cut-off that demarcates which tasks can and cannot be automated is no longer fixed, and machines are free to encroach further into the realm of tasks once performed by workers. But an assumption is still commonly made that, if any entirely new tasks are created by technological change, then these tasks will necessarily be those in which workers, rather than machines, have the comparative advantage (see, for instance, Acemoglu & Restrepo, 2018a, 2018b, 2018c). I will return to look at this assumption more carefully later in this chapter. But intuitively, it provides a significant new source of optimism when thinking about the future of work: though machines might encroach deeper into the realm of *existing* types of tasks performed by workers, if any *new* types of tasks are created then these will necessarily—by assumption—be those better-suited to workers.

Across these newer models, there is also a more explicit recognition that technological progress, and the process of task encroachment that it generates, has two different effects on the labor market. On the one hand, new technologies can allow capital to *substitute* for workers at particular tasks, reducing labor demand for those activities. (As a shorthand, I will talk about "technology" substituting for workers.) But at the same time, they can also *complement* workers at other tasks, increasing the demand for labor at those activities. It follows that the overall effect of technological change on the demand for labor depends on the balance between these two forces: if the complementing force is greater than the substituting force, then the overall demand for the work of human beings increases; but if the substituting force is greater, then that overall demand would instead decline. This distinction between the two forces might sound relatively straightforward. Yet until recently, in both informal commentary on the future of work and the formal economic literature, they were either entangled with one another and hard to distinguish—or simply omitted altogether.

Take the substituting force. This is the effect of technological change that captures the popular imagination today. But in the early formal literature, this effect was often absent: in the canonical model, for instance, there "are no explicitly skill replacing technologies" (Acemoglu & Autor, 2011). The two types of labor are q-complements, an increase in the quantity or productivity of one necessarily increasing the marginal product of the other. As a result, there is no way for technological progress to make either type of labor worse-off. But now, under the task-based approach, where labor is often modeled in competition with capital to perform different tasks, it is possible to identify this substitution force far more clearly.

Or take the complementing force. This is the force that tends to be neglected in popular debate about the future of work. As Autor (2015a) put it, "journalists and expert commentators overstate the extent of machine substitution for human labor and ignore the strong complementarities that increase productivity, raise earnings, and augment demand for skilled labor." Part of the difficulty here is that while the substituting force is easy to imagine—"a worker being replaced by a robot," as it is often put—the complementing force is both much less conspicuous and operates in a variety ways. And the more recent literature is often focused on identifying the different ways these helpful countervailing effects might work.

A final difficulty is that, even when these effects are distinguished from one another in the literature, they are often described in very different ways. For instance, Autor (2015a) uses the simple two-part distinction that I adopt: "machines both substitute for and complement labor." But in Acemoglu and Restrepo (2018c), the substituting force is renamed the "displacement effect" and the complementing force is decomposed into four separate effects: a "productivity effect," "capital accumulation," the "deepening of automation," and a "reinstatement effect." Then in Acemoglu and Restrepo (2019), the substituting force reappears as the "displacement effect," but the complementing force is decomposed into only two different effects: just a "productivity effect" and a "reinstating effect." There is also an additional complication in that the term "automation" is used in different ways as well: Autor (2015a) uses the term automation in the broadest sense, as in this chapter, to refer to the general use of technology in the workplace, whereas Acemoglu and Restrepo (2018c) uses it in a narrower way, to refer to the "expansion in the set of tasks that can be produced with capital" alone (I call this "task encroachment"). None of these approaches are incorrect: all are trying to identify the different ways in which same technological progress might increase or decrease the demand for labor but using different labels in carrying out that exercise.

In any event, understanding this intellectual history, and exploring how economists have changed their minds over time, is of practical use in thinking about a future with insufficient demand for the work of human beings. In early models, such an outcome was often not possible: the complementing force was the focus, and the substituting force was neglected. In later models, this outcome was now possible—but still unlikely. Under the ALM hypothesis, for instance, technology no longer complements all types of workers, but instead only complements those workers who can perform "non-routine" tasks that cannot be automated. In turn, technology substitutes for those workers who perform "routine" tasks that can be automated. But this approach still encourages skepticism about the idea of insufficient demand for human beings because of this large range of "non-routine" tasks that cannot be automated. In short then, while the canonical model supported an optimistic view about the threat of automation by making the strong assumption that technology cannot substitute for labor, early task-based models tended to support the same optimistic view but by making the weaker assumption that "the scope for substitution is bounded" (Autor, 2015c).

However, the scope for substitution has proven not to be bounded in the way that the task-based literature once expected—"non-routine" tasks can increasingly be automated. And in response, recent books, papers, reviews, and reports have tried to work out the new limits of machine capabilities, using a variety of approaches (see Susskind, 2020a for an overview).

But the difficulty with marking out rigid boundaries in this way is that any conclusions one reaches are likely—as with the ALM hypothesis—to swiftly become outdated. A more productive way to think about technological progress is, again, as a process of task encroachment—that machines gradually but relentlessly take on more types of tasks. This involves recognizing that while it is very hard to say what machines might do in the future, it is relatively certain that they will do *more* than they do today. And so, the critical question is whether as new technologies continue this advance, encroaching further into the realm of tasks once performed by human beings, is it possible that the harmful substituting force might eventually overwhelm the helpful complementing force—and the demand for the

work of human beings would decline. This is now a question that, at least in theory, it is possible to explore in the newest "task-based, capability-agnostic" models.

Two Types of Technological Unemployment

It was John Maynard Keynes who first popularized the term "technological unemployment" back in 1930, believing it was a challenge we would "hear a great deal" about in years to come. He thought it would follow from "our discovery of means of economizing the use of labor outrunning the pace at which we can find new uses for labor," but beyond those general fears, he offered few details of how it might happen. In what follows I want build on the intellectual history set out before and distinguish more clearly between two types of technological unemployment, two ways that human beings might find themselves without work due to technological progress. These arguments and ideas are drawn from Susskind (2020a).

Frictional technological unemployment

The first is "frictional" technological unemployment. This is not the sort of technological unemployment that Leontief, whose work was mentioned at the outset, had in mind: here, there is sufficient work for human beings to do. In terms of the two forces, the complementing force continues to overpower the substituting force. The problem is that workers who are displaced from their roles by technology are not able to take up those new roles. Though there are many frictions in the labor market that might create this problem, there are three reasons worth highlighting.

The first friction is the skills mismatch—that workers displaced by new technologies may not have the skills to do the work that has to be done elsewhere in the labor market. Historically, this friction has been an important focus. This is natural, given the empirical evidence from labor markets in the 20th century: that workers with more skills tended to have far better outcomes than those with less skills (i.e., the rising "skill premium"). Economists, for instance, have written of a "race" between technology and education, implying that people have to learn the right skills to keep up (see, for instance, Goldin & Katz, 2008). For that reason, many named the 20th century the "human capital century," capturing the idea that what mattered during that time for an individual's success—and a country's prosperity—was their level of education. And politicians at the time saw "more education" as the most important available response to the challenge of technological change: from U.S. president Bill Clinton hoping in 1996 that "the 13th and 14th years of education [would become] as universal to all Americans as the first 12 are today to U.S. president Barack Obama claiming in 2010 that 'in the coming decades, a high school diploma is not going to be enough. Folks need a college degree. They need workforce training. They need a higher education' " (Susskind, 2020a). All these proposals were attempts to repair the apparent skills mismatch.

But the skills mismatch is not the only important friction. Another is the place-mismatch—that displaced workers may not live in the same geographical location as the work that is

created. From a technological point of view, it is interesting to reflect on how, at the start of the internet era, the rise of new communication technologies led some to predict that these concerns about location would soon no longer matter. This turned out to be a misplaced view—despite "all the hype about the 'death of distance' and the 'flat world,' where you live matters more than ever" (Moretti, 2013). In many ways, this is to be expected. Technological progress is often associated with geographical rise and fall: in the U.S., for instance, the success of Silicon Valley and the decline of the Rust Belt are both regional consequences of technological change (Susskind, 2020a). A growing body of empirical evidence also supports the growing importance of the place-mismatch: Acemoglu and Restrepo (2020a), for instance, show how different local labor markets in the U.S. have vastly different exposure to the use of robots (in large part, because industries differ from place to place); and Autor (2019) argues that the changing nature of jobs in U.S. cities has left non-college urban workers with much less skilled work than they did decades ago (this, he argues, is an important driver of recent declines in non-college wages).

COVID-19 has, in turn, made the importance of place even more apparent. A distinctive feature of the pandemic is that while many white-collar workers have able to retreat to work from home, this has been far less feasible for lower-paid workers in service roles—waiters in restaurants, baristas in coffee shops, retail assistants on the high-street, receptionists in offices—whose roles are far more tightly linked to being physically present in a specific place. One U.S. survey, for instance, found that while 71 percent of people earning more than $180,000 could work remotely during the pandemic, only 41 percent of those earning less than $24,000 were able to do so (Susskind, 2020a).

A third mismatch is the identity-mismatch. Here, it is not that people do not have the right skills or live in the right place—but they have a conception of themselves that is at odds with the available work, and they are willing to stay unemployed to protect that identify. In South Korea, for instance, half of the unemployed are college graduates—and partly, this may be because they are hesitant to take up the low-paid, low-quality roles that they did not believe their education was preparation to do. Or consider adult men of working age in the U.S., displaced from traditional manufacturing roles by automation. There are some that say these men would rather not work at all than take up so-called "pink-collar" work—an unfortunate term, but designed to capture the fact that many of the roles that are hardest to automate are disproportionately done by women; in the U.S., for instance, 92.2 percent of preschool and kindergarten teachers, 92.2 percent of nurses, and 82.5 percent of social workers (Susskind, 2020a).

Structural technological unemployment

The second type of technological unemployment is a structural one. Here, there is simply not enough work for human beings to do, because labor demand declines to the point where workers no longer find it worthwhile to work—even if those frictions from before are resolved. Before exploring this idea, though, it is important to remember that structural technological unemployment is the most extreme outcome that could follow from a decline in labor demand: there are many other possibilities. Indeed, one problem with the term "technological unemployment" itself is that it leads us to think that the only way in which the labor market might respond to a decline in the demand for human beings is through

the number of "jobs" that have to do be done. This is not right: the labor market might adjust in several ways to a decline in demand, not only through the *number* of jobs, but also the *nature* of those jobs—their pay and quality, for instance. To those who find the latter an entirely reasonable proposition but the former hard to fathom, I would encourage them to view the two outcomes as sitting on a continuum of possible consequences, and to recognize that if the demand for the work of human beings were to fall substantially, it is plausible that wages might fall so low for certain people that they prefer not to work at all.

The idea of structural technological unemployment is often dismissed on the grounds that that, ever since modern economic growth began about three centuries ago, people have suffered from bursts of concern about new technologies displacing them from their work—and yet, time and again, they have been wrong. There is thus little evidence, so the skeptical line continues, to support the general fear that technological progress would create large numbers of permanently unemployed workers. And it is true that human beings have been displaced but, time and again, they have been able to take up work elsewhere in the economy. In retrospect, the reason for this repeated mistake can be expressed in terms of the two fundamental forces identified before—over time, in predicting the impact of technological change, people have systematically overestimated the substituting force and underestimated the complementing force. As a result, contrary to their fears, there has always been sufficient demand for the work of human beings.

It follows that, in thinking about the future of work, the key question is whether that favorable balance between those two forces is likely to continue. In my view, an important threat to this balance is the process of task encroachment. And the great merit of the emerging "task-based, capabilities-agnostic" models is that they allow us to explore the impact of this process in a more formal way; as Acemoglu and Restrepo (2018a, 2018c) put it, they allow us to consider the consequences of "the use of machines to substitute for human labor in a widening range of tasks." In what follows, I want to explain why we should take the fear that the balance between these two effects may shift in the future entirely seriously.

In the early literature, as noted, these two effects were either conflated or the substituting force was omitted altogether (i.e., the canonical model). In later literature, these effects were disentangled but the substituting force was still tightly constrained by the assumption that there were firm boundaries to the capabilities of machines (i.e., the ALM hypothesis). In more recent literature, that constraint on the substituting force has been removed (i.e., the newer task-based, capabilities-agnostic models). But there is, in my view, still a residual problem—the process of task encroachment is still considered almost exclusively through its role in strengthening the substituting force and not through how it may weaken the countervailing complementing force as well. Put another way, the process of task encroachment is troubling not only because it widens the range of tasks in which new technologies can substitute for labor but, at the same time, all else held constant, it also narrows the range of tasks in which capital might then complement labor.

The complementing force, weakened

To see how the complementing force might be weakened by the process of task encroachment, consider some of the traditional channels identified in the literature through which technological progress might complement human beings. As noted, the challenge in

identifying these challenges is that economists, over the decades, have broken down this complementing force in different ways. Nevertheless, it is still possible to distinguish three general ways in which this helpful force is thought to work: as a "productivity effect," a "bigger-pie" effect, and a "changing pie" effect (Susskind, 2020a).

To begin with, new technologies might complement workers *directly* by making them more productive at certain tasks. I call this the "productivity effect."[3] When commentators on the future of work describe how new technologies might "help," "augment," "enhance," or "empower" workers, making them more effective at the activities that they do, this is the effect they tend to have in mind. In a traditional economic set-up, this effect would be analogous to factor-augmenting technological progress when using a neoclassical production function. However, the task-based approach allows us to be far more explicit about what exactly workers are becoming more productive at doing—namely, unautomated *tasks*. And so, new technologies might displace people from certain tasks, but they can also make workers more productive at tasks that have not been automated. When those improvements in productivity are passed on to consumers through lower prices or higher-quality goods, that may increase the demand for the work of human beings.[4]

Here, though, is the problem with this effect: future technologies will surely make some human beings more productive at certain tasks, but this will only raise the demand for their efforts if they remain able to perform those tasks more efficiently than machines (i.e., at a lower unit cost). When that is no longer the case, the relative productivity of human beings becomes irrelevant—machines will simply perform the task instead. Take a traditional craft like candle making (Susskind, 2020a). Human beings were once best placed to perform that task—but that ceased to be the case some time ago. The productivity of a present-day tallow-chandler might interest some enthusiasts, but from an economic point of view it is irrelevant—that task is now done by a machine. As the process of task encroachment continues, human capabilities will become irrelevant in this manner for more tasks. In this way, the process not only increases the strength of the substituting force, but it narrows the range of tasks in which the complementing force might directly help human beings through this productivity effect.

But technological progress can also complement human beings *indirectly*. If we think of the economy as a pie, for example, another channel that is widely identified in the literature is the "bigger pie effect."[5] Here, as productivity improves, incomes grow, the demand for goods and services rises, and so demand also increases for all that the tasks that are needed to produce them. In a similar spirit to before, new technologies might displace workers from certain tasks, but a growing demand for unautomated tasks elsewhere in the labor market could provide them with work instead. Consider Autor (2015b):

> I think that people are extremely unduly pessimistic . . . we neglect the fact that as we create wealth, which we certainly do through productivity improvements, we create more consumption. People want more experiences; they want more goods and services. Consequently, as people get wealthier, they tend to consume more, so that also creates demand.

Or Summers (2013), recounting his experience of being a 1970s MIT graduate student:

> The stupid people thought that automation was going to make all the jobs go away and there wasn't going to be any work to do. And the smart people understood that when more was produced, there would be more income and therefore there would be more demand.

The concern with this effect, though, is the same as previously: in the future, incomes are likely to be far greater than today, and demand for goods and services will rise along with them, but this will not necessarily lead to an increase in the demand for the work of human beings. Again, this will only be the case if people remain better placed than machines to perform whatever tasks are newly in-demand (i.e., at a lower unit cost). And as the process of task encroachment continues, it becomes more likely that a machine will be better placed to perform those tasks instead. Once again, this process not only widens the range of tasks in which machines substitute for labor, but it narrows the range of tasks in which new technologies might complement workers (here, indirectly).

Or consider another way in which new technologies might indirectly complement human beings—that technological progress not only makes the economic pie bigger, but it also changes the pie: this is the "changing pie effect." The British economy, for example, is not only more than 100 times the size it was 300 years ago, but its output, and the method of production, has transformed: agriculture now only employs 380,000 people compared to 3.2 million in 1860; manufacturing only employs 40 percent of the number of workers it did back in 1948 (Susskind, 2020a). But these trends have not led to vast pools of unemployed people because the economy is no longer made up of farms and factories alone. As Autor (2015c) puts it so well, there was little chance that "farmers at the turn of the twentieth century could foresee that one hundred years later, healthcare, finance, information technology, consumer electrics, hospitality, leisure and entertainment would employ far more workers than agriculture."

Historically, consumer tastes and preferences have been an important driver of these structural transformations—over time, they not only have larger incomes, but they change how they spend those incomes as well. This means new goods and services are demanded. And so, there may also be a demand for displaced workers to performs tasks that have to be done to produce them. This is popular argument: consider Mokyr et al. (2015), for instance, who writes that "the future will surely bring new products that are currently barely imagined, but will be viewed as necessities by the citizens of 2050 or 2080;" or Autor and Dorn (2013), who claim that technological progress will "generate new products and services that raise national income and increase overall demand for labor in the economy."

In the future, it is entirely possible that human beings will have different wants and needs to today—perhaps currently inconceivable ones. Yet this will not necessarily mean a greater demand for the work of human beings as well. That will only be the case, as with the other helpful effects, if human beings remain best placed to perform whatever tasks have to be done to produce those goods and services. But in the future, as task encroachment continues, machines may simply become the more efficient choice instead. Look at newer parts of economic life, and you can catch a glimpse of how this might unfold. In 1964, the most valuable company in the U.S. was AT&T, which had 758,611 employees. But in 2018 it was Apple, with only 132,000; and in 2019, it was Microsoft with 131,000 (Susskind, 2020a). Neither Apple nor Microsoft existed back in the 1960s. Many of the goods and services would have been hard to imagine in a pre-Internet era.

A common theme runs through the previous discussion of task encroachment and the complementing force—while it is right to recognize that as our economies grow and change the demand for tasks to produce everything will grow and change as well, it is wrong to think that human beings will necessarily be the most efficient choice to perform many of those tasks. I call this the "superiority assumption" (Susskind, 2020a). Economists describe the

demand for labor as a "derived demand," recognizing that workers are only demanded in so far as the goods and services that they produce are demanded. But the task-based approach reveals a deeper sense in which it is a derived demand: workers are only demanded in so far as they are best-placed to do whatever tasks must be done to produce those goods and services. While the superiority assumption holds that might be the case. But as it weakens, it may not.

And so, the argument to take structural unemployment seriously follows: "over time, machines continue to become more capable, taking on tasks that once fell to human beings. The harmful substituting force displaces workers in the familiar way. For a time, the helpful complementing force [working in the ways identified] continues to raise the demand for those displaced workers elsewhere. But as task encroachment goes on, and more and more tasks fall to machines, that helpful force is weakened as well" (Susskind, 2020a). Eventually, the substituting force overruns the complementing force and the demand for the work of human beings falls away. This does not necessarily lead to a technological big bang after which large numbers of people suddenly find themselves without work. Instead, a far more gradual withering in the demand for labor is possible, as the balance between the two forces tips out of favor of human beings. From an empirical point of view, one might argue that this sort of phenomenon can already be seen in the data: Acemoglu and Restrepo (2020a), for instance, studied the use of industrial robotics in the U.S. from 1990 to 2007 and found a relatively recent case of the substituting force overpowering the complementing force: one more robot per 1,000 workers, they observed, reduced the employment-to-population ratio in the U.S. economy by 0.2 percentage points and wages by 0.42 percent. In traditional models of the labor market, it was not possible for new technologies of any types to have these harmful aggregate consequences.

To be clear, this reasoning does not rely on the assumption that machines will be able to do everything in the future. In a recent survey, leading computer scientists made the claim that there is a 50 percent chance that machines will outperform human beings at "every task" within 45 years (Grace et al., 2018). But it is possible that these fears are wrong: tasks might remain that prove impossible to automate, others than are possible but unprofitable to automate, some that remain restricted to human beings for cultural or regulatory reasons, and others that we prefer human beings to do because we value the very fact they are done by a person and not a machine (crafting a fine suit, preparing a delicious meal, caring for one another in ill health and old age). Yet even though machines may not do everything in the future, despite those expert predictions, they will certainly do *more.* And it is this gradual, but relentless, process of task encroachment that is worrisome: human beings forced into a diminishing set of activities, with no economic law to say there must be enough demand for those residual unautomated tasks to keep everyone who wants a job employed at a sufficient wage.

The creation of new tasks

In the new "task-based, capability-agnostic" models, there is a recognition that the complementing force need not be stronger than the substituting force. For instance, in Acemoglu and Restrepo (2018b) a "horse equilibrium" is identified where workers are immiserated by technological progress. This label is of course a reference to the work of

Wassily Leontief, mentioned at the outset, who thought technology would do the same to human beings as it had done in the past to horses. As noted, when Leontief was writing, his view was far from mainstream and very difficult to capture in the core models. But as this more recent work shows, that is no longer the case.

Nevertheless, although Acemoglu and Restrepo (2018b) identify the horse equilibrium as a possible outcome, it is avoided on the balanced growth path in their dynamic model for an intriguing reason—in this model, technological progress is associated with the endogenous creation of "new and more complex" tasks in which labor has the comparative advantage, and so displaced workers can take up these new activities instead. This is another type of helpful changing-pie effect: not only do consumers buy different goods and services, as before, but producers also changed the way in which they make them. Historically, this process appears to have an important source of demand creation (Acemoglu & Restrepo, 2018b, 2018c, 2019). And it is why, Acemoglu and Restrepo (2018b) argue, it is wrong to compare the economic fate of human beings and horses: the former, if displaced, could move on to perform complex new tasks, but the latter would only ever be suited for pulling heavy loads from place to place and could not find anything else to do.

This additional complementing effect might have played an important role in the past. Yet it is not obvious that it is immune from the previous concerns about the effects of task encroachment, either. In Acemoglu and Restrepo (2018b), new tasks are endogenously created because technical change in the model is "directed"—when human beings are displaced, they become cheaper, and so there is an incentive to create new tasks for them to do to take advantage of those lower costs. But we might then ask why this effect did not also help displaced horses find new roles—why was there not also an incentive to create new tasks for them to do? This argument might sound a little playful, but there is a very serious point: the reason that new tasks were not created for horses to do was because their capabilities had been exhausted relative to machines. No matter how low their relative cost, there was no realm of economic activity in which they could be useful re-employed at a similar scale. At the moment, human capabilities are so remarkable relative to machines that it seems entirely reasonable to assume that we can continue to devise more and more new tasks in which labor has the comparative advantage. But as task encroachment continues, and human beings start to resemble horses when compared with machines for more activities, that assumption looks increasingly questionable.

In the most recent literature, there is a growing recognition that this endogenous process of task creation alone may be insufficient to maintain the demand for the work of human beings. There is, for example, a fear that the labor market might tend to engage in "excessive automation" (see, for instance, Acemoglu & Restrepo, 2019, 2020b; Acemoglu et al., 2021).

Rather than passively rely on the endogenous forces identified before, this work involves a call for the state—through taxes and regulation, for instance—to actively strengthen the incentive to develop technologies that complement, rather than substitute, for human beings. This call resembles an increasingly influential movement in computer science, led by Stuart Russell, to incentivize the development of AI that is "provably beneficial" to human beings (see Russell, 2019, for instance). Again, the belief is that AI research, left undirected, will develop in a way that is harmful to human beings, and ought to be steered in a more positive direction. Putting to one side the question of whether calls to redirect technological change are desirable or feasible, the most important observation for the purposes of this chapter is that these interventions are being discussed at all. This demonstrates that the problem of

technological unemployment, where there is not enough demand for the work of human beings to keep everyone is sufficiently well-paid employment, is not simply a new theoretical possibility but one of increasing practical concern, too, that demands our attention.

Taking the Idea Seriously

Earlier in this chapter, I explained that a decline in labor demand need not only manifest in a fall in the quantity of available work, but its quality too—its pay, security, status, and so on. This, in part, explains why those who dismiss technological unemployment as a sudden break from economic life today are likely to be mistaken. As argued in Susskind (2020a), it is not a coincidence that concerns about automation are intensifying at the same time as worries about inequality are growing—the two problems are closely related. Today, the labor market is the main way that we share out income in society; for most people, their job is their primary source of income. The growing inequalities that we see in many labor markets show that this approach is already under stress—some get far more for their efforts than others.

Technological unemployment is but a more radical version of this same story, but one that ends with some workers earning nothing at all. In turn, the argument that this outcome is only a threat if *most* people find themselves without work is an unhelpful distraction: a world where even 15 to 20 percent of people, for example, find themselves without sufficiently well-paid work would still present serious challenges.

If we do take these ideas seriously, what challenges will we have to confront? From an economic point of view, the main problem is a distributional one. In the future, technological progress is likely to make us collectively more prosperous than ever before. The task will be to find a way to share out that prosperity if our traditional way of doing so, paying people for the work that they do, is less effective than in the past. When John Maynard Keynes introduced the prospect of "technological unemployment" he simultaneously dismissed it as a threat—within a century, he thought, we would be so collectively prosperous that the traditional economic problem that haunted our ancestors, the "struggle for subsistence," would be resolved and we could all "live wisely and agreeably and well" (Keynes 1963). His prediction was correct—today, global GDP per head is indeed almost large enough to pull everyone out of poverty. But he also made a big mistake—he assumed that the world's prosperity would automatically be enjoyed by everyone. Already, that is far from the case; the economic pie may be far larger than ever before, but most people's slice of it remains wafer-thin (see Susskind, 2020a and Stiglitz, 2008 for more detail on this argument). And in a world with technological unemployment, this challenge will be even greater.

But technological unemployment will also present us with problems that take us beyond the traditional questions that occupy most economists exploring the impact of technology on work. It is often said, for instance, that work is not simply a source of an income but also of meaning and direction. Understanding this relationship between work and purpose is critical because its nature will necessarily shape the form of any interventions that are required in the future. Suppose, for instance, that for some human beings the link between work and fulfilment is important. For them, a "job guarantee" might be an appealing intervention, providing them with an income as well as meaningful activity. Alternatively,

suppose you look to the almost 70 percent of workers in the U.S. are either "not engaged" in or "actively disengaged" from it, while only 50 percent say they get a sense of identity from their job. For them, a "basic income" might be more appropriate, providing an income but allowing them to find purpose beyond the traditional labor market. These are unfamiliar and challenging ideas. They raise hard questions—not simply how to pay for such a scheme, a preoccupation of many who are concerned with proposals like this, but how to maintain social solidarity in a world where some people do not make an economic contribution to the collective pot through the work that they do.[6] But if the threat of technological unemployment is a real one, then we must now take these new challenges seriously as well.

Acknowledgements

Thank you to Anton Korinek for his expert editorial advice, two anonymous referees for their guidance, and Daron Acemoglu, Katya Klinova, and all other attendees at the online conference for AI Governance in March 2021 for helpful comments and reflections.

Notes

1. The task-based approach has a rich intellectual history (see Susskind & Susskind, 2015 for a detailed discussion). Many classical social theorists, for instance, were interested in the idea, though spoke in terms of the "division of labor." Precursors can be found in the economics literature (Zeira, 1998, for instance, discusses how factors produce "intermediate goods"—rather than "tasks"—which are they combined to produce a unique output). Autor et al. (2003), though, brought it to wider attention in the profession.
2. This system is not without critics. One concern, for instance, raised by several of the authors of the Stanford system, is that it was trained on a data set where some images also contained a ruler. If freckles which are considered suspicious are more likely to measured or marked, then the fear is that this system may actually be partly detecting rulers rather than cancers (or, as one study put it, learning "that rulers are malignant," see Narla et al., 2018). How serious is this problem? It is important to note that, despite this bias, the original system still appeared to outperform human doctors in identifying cancers in unseen photos of freckles. In any event, as several of the authors of the Stanford system have noted, exploring the biases associated with non-standardized images is an important task when designing these systems (see, for instance, Esteva & Topol, 2019).
3. Acemoglu and Restrepo (2018c) and Acemoglu and Restrepo (2019) also talk of the "productivity effect" in their framework.
4. This is analogous to the idea in Autor (2015c), that "workers are more likely to benefit directly from automation if they supply tasks that are complemented by automation."
5. If the "productivity effect" is the partial equilibrium effect of an increase in productivity—workers become more productive at their work, and their wages might rise as a result—the "bigger pie effect" can be thought of as a general equilibrium effect of an increase in productivity.
6. For that reason, for instance, some might prefer a "conditional" basic income, where strings are attached to any financial support, rather than a "universal" basic income, which is given without any conditions. This distinction is explored at greater length in Susskind (2020a) and Susskind (2021).

References

Acemoglu, Daron. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature 40* (1), 7–72.

Acemoglu, Daron, & Autor, David. (2011). Skills, tasks and technologies: Implications for employment and earnings. In David Card and Orley Ashenfelter (Eds.), *Handbook of labour economics, volume 4B* (pp. 1043–1171). North-Holland.

Acemoglu, Daron, & Autor, David. (2012). What does human capital do? A review of Goldin and Katz's "The race between education and technology". *Journal of Economic Literature 50* (2), 462–463.

Acemoglu, Daron, & Restrepo, Pascual. (2018a). Modelling automation. *AEA Papers and Proceedings 108* , 48–53.

Acemoglu, Daron, & Restrepo, Pascual. (2018b). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review 108* (6), 1488–1542.

Acemoglu, Daron, & Restrepo, Pascual. (2018c). Artificial intelligence, automation, and work. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda*, 197–236. University of Chicago Press.

Acemoglu, Daron, & Restrepo, Pascual. (2019). Automation and new tasks: How technology displaces and reinstates labour. *Journal of Economic Perspectives 33* (2), 3–33.

Acemoglu, Daron, & Restrepo, Pascual. (2020a). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy 128* (6).

Acemoglu, Daron, & Restrepo, Pascual. (2020b). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society 13* (1), 25–35.

Acemoglu, Daron, Autor, David, Hazell, Jonathan, & Restrepo, Pascual. (2021). AI and jobs: Evidence from online vacancies. NBER Working Paper Series No. 28257.

Aghion, Philippe, Jones, Benjamin, & Jones, Charles. (2019). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Autor, David. (2013). The "task approach" to labor markets: An overview. *Journal for Labour Market Research 46* (3), 185–199.

Autor, David. (2015a). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives 29* (3), 3–30.

Autor, David. (2015b). The limits of the digital revolution: Why our washing machines won't go to the moon. An interview with SocialEurope.eu. https://www.socialeu-rope.eu/2015/10/the-limits-of-the-digital-revolution-why-our-washing-machines-wont-go-to-the-moon/. October.

Autor, David. (2015c). Polanyi's paradox and the shape of employment growth. Re-evaluating Labor Market Dynamics: A Symposium Sponsored by the Federal Reserve Bank of Kansa City, 129–179, Jackson Hole, Wyoming, August 21–23, 2014.

Autor, David. (2019). Work of the past, work of the future. Richard T. Ely Lecture, *AEA Papers and Proceedings 109* , 1–32.

Autor, David, & Dorn, David. (2013). Technology anxiety: Past and present. Bureau for Employers' Activities, International Labour Office, December.

Autor, David, Katz, Lawrence, & Kearney, Melissa. (2006). The polarisation of the U.S. labor market. *The American Economic Review 96* (2), 189–194.

Autor, David, Katz, Lawrence, & Kearney, Melissa. (2008). Trends in U.S. wage inequality: Revising the revisionists. *The Review of Economics and Statistics 90* (2), 300–323.

Autor, David, Katz, Lawrence, & Krueger, Allan. (1998). Computing inequality: Have computers changed the labor market? *The Quarterly Journal of Economics 113* (4), 1169–1214.

Autor, David, Levy, Frank, & Murnane, Richard. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics 118* (4), 1279–1333.

Bekman, Eli, Bound, John, & Machin, Stephen. (1998). Implications of skill-biased technological change: International evidence. *The Quarterly Journal of Economics 113* (4), 1245–1279.

Bound, John, & Johnson, George. (1992). Changes in the structure of wages in the 1980's: An evaluation of alternative explanations. *The American Economic Review 82* (3), 371–392.

Card, David, & Lemieux, Thomas. (2001). Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *The Quarterly Journal of Economics 116* (2), 705–746.

Esteva, Andre, & Topol, Eric. (2019). Can skin cancer diagnosis be transformed by AI? *The Lancet 394* (10211), 1795.

Esteva, Andre, Kuprel, Brett, Novoa, Roberto A., Ko, Justin, Swetter, Susan M., Blau, Helen M., & Thrun, Sebastian. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542* , 115–118.

Goldin, Claudia, & Katz, Lawrence. (1998). The origins of technology-skill complementarity. *The Quarterly Journal of Economics 113* (3), 693–732.

Goldin, Claudia, & Katz, Lawrence. (2008). *The race between education and technology*. Harvard University Press.

Goos, Maarten, & Manning, Alan. (2007). Lousy and lovely jobs: The rising polarisation of work in Britain. *The Review of Economics and Statistics 89* (1), 118–133.

Goos, Maarten, Manning, Alan, & Salomons, Anna. (2009). Job polarisation in Europe. *The American Economic Review 99* (2), 58–63.

Goos, Maarten, Manning, Alan, & Salomons, Anna. (2014). Explaining job polarization: Routine-biased technological change and offshoring. *The American Economic Review 104* (8), 2509–2526.

Grace, Katja, Salvatier, John, Dafoe, Allan, Zhang, Baobao, & Evans, Owain. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research 62* , 729–754.

Harrod, Roy. (1948). *Towards a dynamic economics*. Macmillan.

Hayes-Roth, Frederick, Waterman, Donald, & Lenat, Douglas. (1983). *Building expert systems*. Addison-Wesley.

Hicks, John. (1932). *The theory of wages*. Macmillan.

Katz, Lawrence, & Murphy, Kevin. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *The Quarterly Journal of Economics 107* (1), 34–78.

Keynes, John Maynard. (1963). *Essays in persuasion*. W. W. Norton.

Leontief, Wassily. (1979). Is technological unemployment inevitable? *Challenge* 22(4), 48–50.

Leontief, Wassily. (1983a). National perspective: The definition of problems and opportunities. In *The Long-term Impact of Technology on Employment and Unemployment: A National Academy of Engineering Symposium, 30 June 1983*. National Academy Press.

Leontief, Wassily. (1983b). Technological advance, economic growth, and the distribution of income. *Population and Development Review* 9(3), 403–10.

Levy, Frank, & Murnane, Richard. (2004). *The new division of labor*. New Age International Ltd.

Mokyr, Joel, Vickers, Chris, & Ziebarth, Nicholas. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives 29* (3), 31–50.

Moll, Benjamin, Lukasz, Rachel, & Restrepo, Pascual. (2021). Uneven growth: Automation's impact on income and wealth inequality. NBER Working Paper No. 28440.

Moretti, Enrico. 2013. *The new geography of jobs*. Mariner Books.

Narla, Akhila, Kuprel, Brett, Sarin, Kavita, Novoa, Roberto, & Ko, Justin. (2018). Automated classification of skin lesions: From pixels to practice. *Journal of Investigative Dermatology 138* (10), 2108–2110.

Polanyi, Michael, (1966). *The Tacit Dimension*. Chicago University Press.

Solow, Robert. (1956). A contribution to the theory of economic growth. *The Quarterly Journal of Economics 70* (1): 65–94.

Stiglitz, Joseph. (2008). Toward a general theory of consumerism: Reflections on Keynes's economic possibilities for our grandchildren. In Lorenzo Pecchi & Gustavo Piga (Eds.), *Revisiting Keynes: Economic possibilities for our grandchildren*, 41–85. MIT Press.

Summers, Lawrence. (2013). Economic possibilities for our children. The 2013 Martin Feldstein Lecture. NBER Reporter, No. 4.

Susskind, Daniel. (2016). Technology and employment: Tasks, capabilities, and tastes. Thesis for DPhil in Economics, Oxford University.

Susskind, Daniel. (2019). Re-thinking the capabilities of technology in economics. *Economics Bulletin 39* (1), 280–288.

Susskind, Daniel. (2020a.) *A world without work: Technology, automation, and how to respond*. Allen Lane.

Susskind, Daniel. (2020b). Work in the digital economy. In Robert Skidelsky and Nan Craig (Eds.), *Work in the future*, 125–132. Palgrave MacMillian.

Susskind, Daniel. (2020c). A model of task encroachment in the labour market. Oxford University Working Paper.

Susskind, Daniel. (2021). A world with less work. A Boston Review Forum on Redesigning AI, (May).

Susskind, Daniel, & Susskind, Richard. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press.

Russell, Stuart. (2019). *Machine compatible*. Allen Lane.

Russell, Stuart, & Norvig, Peter. (1995). *Artificial intelligence: A modern approach*. Pearson Education.

Winston, Patrick. (1977). *Artificial intelligence*. Addison Wesley.

Zeira, Joseph. (1998). Workers, machines and economic growth. *The Quarterly Journal of Economics 113* (4), 1091–1113.

CHAPTER 34

HARMS OF AI

DARON ACEMOGLU

Introduction

To many commentators, artificial intelligence (AI) is the most exciting technology of our age, promising the development of "intelligent machines" that can surpass humans in various tasks; create new products and services; or even build other machines that can improve themselves, perhaps beyond all human capabilities. The last decade has witnessed rapid progress in AI, based on the application of modern machine learning techniques and huge amounts of computational power to massive, often unstructured data sets (e.g., Russell & Norvig, 2009; Neapolitan & Jiang, 2018; Russell, 2019).[1] AI algorithms are now used by almost all online platforms and in industries that range from manufacturing to health, finance, wholesale, and retail (e.g., Ford, 2015; Agarwal et al., 2018; West, 2018). Government agencies have also started relying on AI, especially in the criminal justice system and in customs and immigration control (e.g., Thompson, 2019; Simonite, 2020).

Whether AI will be everything its enthusiastic creators and boosters dream or not, it is likely to have transformative effects on the economy, society, and politics in the decades to come. Some of these effects are already visible in AI algorithms' impact on social media, data markets, monitoring of workers, and work automation. Like many technological platforms (or "general purpose technologies") that can be used for the development of a variety of new products, services, and production techniques, there are a lot of choices about how AI technologies will be developed. This, combined with the pervasive effects of AI throughout society, makes it particularly important that we consider its potential dark side as well.

In this chapter, I will focus on three broad areas in which the deployment of AI technologies may have economic and social costs if not properly regulated. I want to emphasize at the outset that the arguments I will present are *theoretical*—currently, there is insufficient empirical evidence to determine whether the mechanisms I isolate are important in practice. The spirit of the exercise is to understand the *potential harms* that unregulated AI may create so that we have a better understanding of how we should track and regulate its progress.

The areas I will focus on are:

Collection and control of information

I will argue that the combination of the demand from AI technologies for data and the ability of AI techniques to process vast amounts of data about users, consumers, and citizens produces a number of potentially troubling downsides. These include: (a) *privacy violation:* companies and platforms may collect and deploy excessive amounts of information about individuals, enabling them to capture more of the consumer surplus via price discrimination or violate their privacy in processing and using their data; (b) *unfair competition*: companies with more data may gain a strong advantage relative to their competitors, which both enables them to exercise market power to extract surplus and also relaxes price competition in the marketplace, with potentially deleterious effects; and (c) *behavioral manipulation:* data and sophisticated machine learning techniques may enable companies to identify and exploit biases and vulnerabilities that consumers themselves do not recognize, thus pushing consumers to lower levels of utility and simultaneously distorting the composition of products in the market.

Labor market effects of AI

I will argue that, even before AI, there was too much investment in cutting labor costs and wages in the United States (and arguably in other advanced economies as well). Such efforts may be excessive either because, in attempting to cut costs, they reduce production efficiency, or because they create non-market effects (for example, on workers losing their jobs or being forced to take lower-pay work). In principle, as a broad technological platform, AI could have rectified this trend—for example, by promoting the creation of new labor-intensive tasks or by providing tools for workers to have greater initiative. This does not seem to have taken place. Instead, automation is a quintessential example of efforts to cut labor costs, and like other efforts, it may be excessive. Many current uses of AI involve automation of work or the deployment of AI in order to improve monitoring and keep wages low, motivating my belief that AI may be exacerbating the excessive efforts to reduce labor costs. In this domain, I focus on four broad areas: (a) *automation*: I explain why automation, a powerful way to reduce labor costs, can be part of the natural growth process of an economy, but it can also be excessive because firms do not take into account the negative impact of automation on workers; (b) *composition of technology*: problems of excessive automation intensify when firms have a choice between investing in automation versus new tasks—I explain why this margin of technology choice may be severely distorted and how AI technologies may further distort this composition; (c) *loss of economies of scope in human judgment*: in contrast to the hope that AI will take over routine tasks and, in the process, enable humans to specialize in problem-solving and creative tasks, AI-human interplay might gradually turn humans into worse decision-makers as they hand over more and more decisions to machines, especially when there are economies of scope across tasks; and (d) *monitoring*: I also explain how technologies like AI that increase the monitoring ability of employers are very attractive to firms, but may at the same time generate significant social inefficiencies.

AI, communication, and democracy

Finally, I will suggest that AI has also exacerbated various political and social problems related to communication, persuasion, and democratic politics that once again predate the onset of this technology. The main concern here is that democratic politics may have become more difficult, or even fundamentally flawed, under the shadow of AI. I focus on: (a) *echo chambers in social media*: how AI-powered social media generates echo chambers that propagate false information and polarize society; (b) *problems of online communication*: online social media, which is interwoven with AI, may disadvantage fruitful political exchange relative to offline communication because there is less accumulated trust between communicators, which favors non-political speech and "broadcasting" rather than bilateral informative communication; (c) *big brother effects*: AI increases the ability of governments to closely monitor and stamp out dissent; and (d) *automation and democracy:* finally, I suggest that the process of automation may further damage democracy by making workers less powerful and less indispensable in workplaces. Each of these outcomes is damaging, but their most consequential effects are mediated by imparing democratic discourse.

In each one of the above instances, I discuss the basic ideas about potential costs, their root causes and how they will exhibit themselves in practice. Throughout, my approach will be informal, attempting to communicate the main ideas. For this reason, I present some of the details of the models that clarify these mechanisms in the Appendix. Some of these models are based on existing work and others are offered as ideas for future exploration.

In addition, I will point out the relevant context and evidence in some cases, even though, as already noted, we do not have sufficient evidence to judge whether most of the mechanisms I explore here are likely to be important in practice. I then discuss the common aspects of the potential harms from AI and explore their common roots. I also argue that these costs, if proved important, cannot be avoided in an unregulated market. In fact, I will suggest that in many of these instances, greater competition may exacerbate the problem rather than resolving it.

The aforementioned list leaves out several other concerns that experts have expressed (e.g., AI leading to "misaligned" super intelligence or AI's effects on war and violence), because of space restrictions and because they are further away from my area of expertise. I mention them briefly toward the end of the chapter.

The long list of mechanisms via which AI could have negative economic, political, and social effects may create the impression that this technological platform is *bound* to have disastrous social consequences, or it may suggest that some of these problems are solely created by AI. Neither is true. Nor am I particularly opposed to this technology. I believe that AI is a hugely promising technological platform. Furthermore, with or without AI, our society has deep problems related to the power of corporations, automation and labor relations, and polarization and democracy. AI exacerbates these problems *because* it is a powerful technology and, owing to its general-purpose nature and ambition, it is applicable in a wide array of industries and domains, which amplifies its ability to deepen existing fault lines. These qualities make the potential negative effects of AI quite difficult to foresee as well. Perhaps even more than with other technologies and platforms, there are many directions—with hugely different consequences—in which AI can be developed. This variability makes it doubly important to consider the costs that AI *might* create. It is also vital to think about the *direction* of development for this technology.

Indeed, my point throughout is that AI's costs are avoidable. If they were to transpire, this would be because of the choices made and the direction of research pursued by AI developers and tech companies. They would also be due to the lack of appropriate regulation by government agencies and societal pressure to discourage nefarious uses of the technology and to redirect research away from them. This last point is important: again, like most other technologies, but only more so, the direction of research of AI will have major distributional consequences and far-ranging implications for power, politics, and social status. It would be naïve to expect that unregulated markets would make the right trade-offs about these outcomes—especially because, at the moment, major decisions about the future of AI are being made by a very small group of top executives and engineers in a handful of companies. Put differently, AI's harms are the harms of unregulated AI. But in order to understand what needs to be regulated and what the socially optimal choices may be, we first need to systematically study the potential downside of this technology. It is in this spirit that the following essay is written.

The rest of this essay is organized as follows. To begin, I start with the effects that AI creates via the control of information. The chapter then moves to discuss AI's labor market implications. I then turn to the effects of this technological platform on social communication, polarization and democratic politics. Next, the chapter briefly touches upon a few other potential unintended consequences of AI technologies. I will then step back and discuss the role of choice in this process. I explain why the direction of technological change in general, and the direction of AI research in particular, is vital, and how we should think about it. This discussion also reiterates that many of the costs are the result of choices made about the development and use of AI technologies in specific directions. I then build on the mechanisms discussed in the chapter to emphasize that unregulated markets are unlikely to internalize AI's costs, that greater competition may sometimes make things worse, and that unfettered markets are unlikely to direct technological change towards higher social-value uses of AI. In this spirit, I also provide some ideas about how to regulate the use of AI and the direction of AI research. The Appendix contains the details of the models discussed informally in the text.

AI and Control of Information

Data is the lifeblood of AI. The currently-dominant approach in this area is based on turning decision problems into prediction tasks and applying machine learning tools to very large data sets in order to perform these tasks. Hence, most AI researchers and economists working on AI and related technologies start from the premise that data creates positive effects on prediction, product design, and innovation (e.g., Brynjolfsson & McAfee, 2014; Jones & Tonetti, 2020; Varian, 2009; Farboodi et al., 2019). However, as emphasized by several legal scholars and social scientists, data and information can be misused—deployed in exploitative ways that benefit digital platforms and tech companies at the expense of consumers and workers (e.g., Pasquale, 2015; Zuboff, 2019). Zuboff, for example, argues that such exploitative use of data is at the root of the recent growth of the tech industry, which "claims human experience as free raw material for hidden commercial practices of extraction, prediction, and sales." (2019, p. 8).

First we must discuss the social costs of AI related to the control of data and information, with a special emphasis on exploring when data can become a tool for excessive extraction and prediction.

Too much data

Concerns about control and misuse of information become particularly important when there are benefits to individuals from "privacy". Individuals may value privacy for instrumental or intrinsic reasons. The former includes their ability to enjoy greater consumer surplus, which might be threatened if companies know more about their valuations and can charge them higher prices. The latter includes various characteristics and behaviors that individuals would prefer not to reveal to others. This could be for reasons that are economic (e.g., to avoid targeted ads), psychological (e.g., to maintain a degree of autonomy), social (e.g., to conceal certain behaviors from acquaintances), or political (e.g., to avoid persecution).

Standard economic analyses tend to view these privacy-related costs as second-order for two related reasons: if individuals are rational and are given decision rights, then they will only allow their data to be used when they are compensated for it adequately, and this would ensure that their data will be used by companies only when the benefits exceed the privacy costs (e.g., Varian, 2009; Jones & Tonetti, 2020). Secondly, in surveys, individuals appear to be willing to pay only a little to protect their privacy, and hence the costs may be much smaller than the benefits of data (e.g., Athey et al., 2017). Yet these arguments have limited bite when the control of data has a "social" dimension—meaning that when an individual shares her data, she is also providing information about others. This social dimension is present, by default, in almost all applications of AI because the use of data is specifically targeted at learning from like-cases in order to generalize and apply the lessons to other settings.

How does this social dimension of data affect the costs and benefits of data? This question is tackled in a series of papers, including MacCarthy (2011), Choi et al. (2019), Bergemann et al. (2021), and, serving as the basis of this discussion, Acemoglu, Makhdoumi et al. (2022). The social dimension of data introduces two interrelated effects. First, there will be *data externalities*: when an individual shares her data, she reveals information about others. To the extent that data is socially valuable and individuals do not internalize this, data externalities could be positive. But if indirect data revelation impacts the privacy of other individuals, these externalities could be negative. The second effect is what Acemoglu, Makhdoumi et al. call *submodularity*: when an individual shares her data and reveals information about others, this reduces the value of others' information both to themselves and to potential data buyers (such as platforms or AI companies). This is for the simple reason that when more information is shared about an individual, the individual's own data becomes less important for predicting his or her decisions.

The Appendix provides a bare-bones model useful for understanding the role of data externalities and submodularity. Here I communicate the main ideas by discussing two simple examples (see the Appendix for details). Consider a platform with two users and suppose that one of them, user 1, has a relatively low value of privacy, while the other, user 2, has a high value of privacy. Data externalities are rooted in the fact that the two users

have data that are very highly correlated, so the platform can learn a lot about user 2 from user 1's data. Given the low value that user 1 attaches to her privacy, the platform will always purchase her data. But this also implies that it will indirectly learn about user 2 given the correlation between the two users' data. If user 2's value of privacy is sufficiently large, it would be socially optimal to shut down data transactions and not allow user 1 to sell her data either. This is because she is indirectly revealing information about user 2, whose value of privacy is very large. This illustrates how data externalities lead to inefficiency. In fact, if user 2's value of privacy is very large, the equilibrium, which always involves user 1 selling her data, can be arbitrarily inefficient.

More interesting are the consequences of submodularity, which can be illustrated using the same example as well. To understand these, let us consider the edge case where the information of the two users is very highly correlated, so that the platform can learn everything relevant about user 2 from user 1's data, or vice versa. The important observation is that this data leakage about user 2 undermines the willingness of user 2 to protect her data. In fact, since user 1 is revealing almost everything about her, she would be willing to sell her own data for a very low price. In this extreme case with very highly correlated data, therefore, both the willingness of the platform to buy user 2's data and benefits user 2 receives from protecting her data are very small, and thus this price becomes approximately zero. But here comes the disturbing part for data prices and the functioning of the data market in this instance: once the second user is selling her data, this also reveals the first user's data almost perfectly, so the first user can only charge a very low price for her data as well. As a result, the platform will be able to acquire both users' data at approximately zero price. This price, obviously, does not reflect users' value of privacy. They may both wish to protect their data and derive significant value from privacy. Nevertheless, the market will induce them to sell their data for close to zero price. Imagine once again that user 2's value of data is also very high. Then, despite this high value of privacy to one of the users, there will be a lot of data transactions, data prices will be near zero, and the equilibrium will be arbitrarily inefficient. These consequences follow from submodularity.

As a second example, consider the case in which both users have the same value of privacy and suppose that these are sufficiently high that the platform could not fully compensate them and still make profits from acquiring their data. However, if the two users' data are sufficiently correlated, the platform may still be able to acquire their data, once again because of submodularity. The reasoning is similar to what we encountered in the previous example: when user 1 is expected to sell her data, this reduces the value of user 2's data, such that user 2 would rather cheaply sell her own data than be uncompensated for the platform's tacit use. Likewise, when user 1 expects user 2 to sell her data, user 1 would do likewise. This "coordination failure" between the two users is created and exploited by the platform to generate value but, critically, this is at the expense of the users. It is straightforward to see that this outcome is a consequence of submodularity as well: when each user expects the other one to sell their data, they become less willing to protect their own data and more willing to sell it relatively cheaply. This locks both users into an equilibrium in which their data is less valuable than they would normally assume and as a result there is, again, too much data transaction.

One final conclusion is worth noting. In addition to the results of excessive data transactions and use, these externalities also shift the distribution of surplus in favor of the platform. Even when it is socially optimal for data to be shared with and used by the

platform, data externalities and correlation between the data of the two users will mean that economic benefits from data will be captured by platforms—not by the users. In particular, when the two users' data is very highly correlated, data prices will again be driven to zero, and thus all of the benefits from the use of data will be captured by the platform.

Are data externalities and the inefficiencies they create empirically relevant? Like many of the channels I discuss in this essay, the answer is that we do not know for sure. If, as industry insiders presume, benefits of data are very large, then they will outweigh the costs from data externalities I have highlighted here. Even in this case, the market equilibrium will not be fully efficient due to disincentives of platforms to compensate users nontrivially for their data, though the use of data by platforms and corporations may be welfare-increasing overall. However, there are reasons to believe that privacy considerations may be quite important in practice. First, many digital platforms have a monopoly or quasi-monopoly situation (such as Google, Meta, or Amazon), and thus their ability to extract rents from consumers can be significant. Second, some of the intrinsic reasons for consumers to care about privacy—related to dissent and civil society organizing—are becoming more important, as I discuss further in this chapter.

In summary, the general lessons in this case are clear: when an individual's data is relevant to others' behavior or preferences (which is the default case in almost all applications of data), then there are new economic forces we have to take into account, and these can create costs from the use of data-intensive AI technologies. In particular:

1. The social nature of data—enabling companies to use an individual's data for predicting others' behavior or preferences—creates externalities, which can be positive or negative. When negative externalities are important, corporations and platforms will tend to overuse data.
2. The social nature of data additionally generates a new type of submodularity, which makes each individual less willing to protect their own data when others share theirs. This submodularity adds to the negative externalities, but even more importantly, it implies that data prices will be depressed and will not reflect users' true values of data and privacy.
3. In addition to the result of excessive data use, both of these economic forces have first-order distributional consequences: they shift surplus from users to platforms and companies.

If these costs of data use and AI are important, they also call for regulating data markets. Some regulatory solutions are discussed in Acemoglu, Makhdoumi et al. (2022), and I return to a more-general discussion of regulation of AI technologies and data toward the end of this chapter.

Data and unfair competition

AI technologies amplify the ability of digital platforms and companies to use data to predict consumer preferences and behavior. On the upside, this might enable firms to design better products for customers (after all, this is one of the main benefits of AI). But the use of such data can also change the nature of competition. These effects become even more

pronounced when some firms are much better placed to collect and use data relative to their competitors, and this is the case I will focus on this subsection. Specifically, one firm's collection and use of data that others cannot access may create a type of "unfair competition", enabling this firm to capture much of the consumer surplus in the market and relax price competition.

A model elucidating how this may happen is provided in the Appendix. Here I will provide the high-level overview. In the pre-AI environment, two firms compete to attract customers. This forces both to charge relatively low prices, enabling consumers to capture most of the benefits from this market.

Next consider the post-AI environment in which one of the firms, say firm 1, can use machine learning, big data, and other methods in order to learn the exact preferences of the consumers it is supplying. This has an economic benefit: firm 1 can customize its products for each consumer. But it also has a dark side: because firm 1 now knows much more about its customers' exact preferences and willingness to pay, it can now price these products so as to capture all or most of the consumer surplus. As a result, the direct effect of the use of AI technologies by firm 1 will be a transfer of economic value from consumers to the firm.

There is an indirect effect, potentially harming consumers as well. If firm 1 now focuses on providing customized products to its consumers, it may cease competing for the rest of the consumers (for whom it does not have detailed data). This will relax competition in the market, enabling the other firm, say firm 0, to also charge higher prices and thus capture some of the surplus previously obtained by consumers.

These two effects, taken together, tend to harm consumers and benefit firms—especially firms that can use AI technology to collect and process extensive data about their customer bases.

In summary, the general lessons from this discussion are complementary to the ones from the previous subsection:

1. The use of AI technologies and detailed consumer data for prediction may improve the ability of firms to customize products for consumers, potentially improving overall surplus.
2. However, it also increases the power of these companies over consumers.
3. This has direct distributional implications, enabling AI-intensive firms to capture more of the consumer surplus.
4. The indirect effect of the better collection and processing of data by one firm is to relax price competition in the market, increasing prices and amplifying the direct distributional effects.

Although in this model the overall surplus in the economy increases after the introduction of AI technologies, in the previous discussion we saw that this is not necessarily true in the presence of other data-related externalities. We will next encounter a new economic force distorting the composition of products offered by platforms.

Behavioral manipulation

The previous section discussed how even the beneficial use of improved prediction about consumer preferences and behavior might have a downside. But improved prediction tools

can also be put to nefarious uses, with potentially far-ranging negative effects. Platforms that collect and effectively process huge amounts of data might able to predict consumer behavior and biases beyond what the consumers are aware of themselves. Anecdotal examples of this concern abound. They include the chain store Target successfully forecasting whether women are pregnant and sending them hidden ads for baby products, or various companies estimating "prime vulnerability moments" and sending ads for products that tend to be purchased impulsively during such moments. They also include marketing strategies targeted at "vulnerable populations", such as the elderly or children. Less extreme advertising strategies also have elements of the same type of manipulation: for example, when websites favor products such as credit cards or subscription programs with delayed costs and short-term benefits or when companies like YouTube and Meta use their algorithms to individuate, favoring more-addictive videos or news feeds based on each user's preferences. As legal scholars Hanson and Kysar have noted, "Once one accepts that individuals systematically behave in nonrational ways, it follows from an economic perspective that others will exploit those tendencies for gain." (1999, p. 630).

Although these concerns are as old as advertising itself, economists and policy-makers hope that consumers will learn how to shield themselves against abusive practices. The sudden explosion in the capabilities of digital platforms to use AI technologies and massive datasets to improve their predictions, however, undercuts this argument. Learning dynamics that had made consumers well-adapted to existing practices would be quickly outdated in the age of AI and big data. Throughout this chapter, I discuss the main issues informally, though they are formally developed in Acemoglu, Makhdoumi, et al. (2023) and I provide an exposition in the Appendix.

Suppose that a platform can find out more about what types of products a consumer is likely to enjoy, including which products she would be tempted to consume in the short run. In a world where the platform has interests aligned with those of the consumer, this additional information may improve welfare. The platform could nudge the consumer (either via which options it presents to her or through its pricing) to purchase the products that she is likely to enjoy. This is the optimistic take about how platforms could use their massive informational advantage in a world of AI and abundant data.

However, in most realistic scenarios, the interests of the platform and the consumers are likely to be misaligned. Consider a simple example. Suppose that the consumer in question may have some "tendencies" or biases that can be exploited in the short run—for example, she may be tempted to purchase products that provide instant gratification or have longer-term costs. From the viewpoint of overall welfare, it would be better for the consumer not to purchase these products, which ultimately might be very costly for her. And yet, if the platform can forecast which products create such temptations for the consumer, it will have a profit incentive to steer her towards these products. Thus, the greater informational advantage of the platform may lead to a type of "behavioral manipulation", meaning that the platform will manipulate the behavior of the consumer in a direction that is deleterious to consumer welfare.

It is also worth noting that, in contrast to the pattern in the previous subsection, this type of behavioral manipulation would not only increase platform profits at the expense of consumers, but it would also distort consumption as it lures consumers towards lower-quality products, reducing overall welfare.

The general lessons in this case are complementary to, but different than, the ones I previously highlighted:

1. AI technologies can enable platforms to know more about consumers' preferences than the consumers do, themselves.
2. This creates the potential for behavioral manipulation, whereby the platform can extract more surplus from consumers while distorting their choices.
3. This type of behavioral manipulation tends to do more than just shift surplus from consumers to the platform; it also distorts the composition of consumption, creating new inefficiencies.

Labor Market Effects of AI

US labor markets have not been doing well for workers over the last 40 years. Wage growth since the late 1970s has been much slower than during the previous three decades, while the share of capital in national income has grown significantly (Acemoglu & Autor, 2011; Autor, 2014). Additionally, wage growth, such as it is, has been anything but shared. While wages for workers at the very top of the income distribution—those in the highest tenth percentile of earnings or those with postgraduate degrees—have continued to grow, workers with a high school diploma or less have seen their real earnings fall. Even college graduates have gone through lengthy periods of little real wage growth.

Many factors have contributed to this sluggish average wage growth and real wage declines at the bottom of the distribution. The erosion of the real value of the minimum wage, which has fallen by more than 30 percent since 1968, has been clearly important for low-wage workers (Lee, 1999). The decline in the power of trade unions and much of the private sector may have played a role as well. The enormous increase in trade with China also likely contributed, by forcing the closure of many businesses and large job losses in low-tech manufacturing industries such as textiles, apparel, furniture, and toys (Autor et al., 2013).

My own work with Pascual Restrepo (Acemoglu & Restrepo, 2019, 2021) emphasizes and documents the importance of the direction of technological progress in this process. In the four decades after World War II, automation went hand-in-hand with new task creation, which contributed to labor demand. A very different path of technological development emerged starting in the 1980s, however—this one exhibiting more automation and fewer advances in worker-friendly technologies to create new tasks (Acemoglu & Restrepo, 2019). Automation eliminated routine tasks in clerical occupations and on factory floors, depressing the demand and wages of workers specializing in blue-collar jobs and clerical functions. Meanwhile professionals in managerial, engineering, financial, designing, and consulting occupations flourished—both because they were essential to the success of new technologies and because they benefited from the automation of tasks that complemented their own work. As automation gathered pace, the wage gap between the top and the bottom of the income distribution magnified. In Acemoglu and Restrepo (2021), we estimate that automation has been, possibly, the most important factor in reshaping the US

wage structure, explaining approximately 50 to 70 percent of the variance of changes in wages between 1980 and 2016.

All of this predates AI. In Acemoglu, Autor et al. (2021), we find that AI activity across US establishments picks up speed only after 2016. Nevertheless, this background is useful because AI may be the next phase of automation. There is evidence that it is already being used both for automation and for tighter monitoring of workers, thus further depressing wages and the labor share of income. In this section, I first explain how automation works and why we may be concerned about excessive automation in general, and how AI may exacerbate these concerns. I then discuss how AI could be used to generate new tasks and technologies that complement humans, but whether this will be the case or not depends on technology adoption and the research and development choices of companies. In this context, I suggest reasons for being concerned that the composition of AI research may be heavily distorted. I also discuss why the most benign view of AI's role in the labor market—automating routine jobs, so that workers have time for more creative, problem-solving tasks—may need to be qualified. Finally, I explore how AI may have pernicious effects when it is used for monitoring.

Excessive Automation and AI

In order to situate the role of AI in the broader context of automation technologies, I start with an informal review of the framework from Acemoglu and Restrepo (2018, 2019), which models the automation of tasks (as well as the creation of new tasks).

In this framework, production is modeled as the completion of a range of tasks. For example, textile production requires cleaning, carding, combing, and spinning the fibers; weaving, knitting, and bonding the yarn; designing, dying, chemical processing, and finishing the textile good; and various non-production tasks, including marketing, advertising, transport, wholesale, retail, and so on.

The key economic decision centers on the allocation of these tasks to capital and labor (and to different types of labor). Automation corresponds to expanding the set of tasks performed by capital. In addition to automation and other types of technologies, Acemoglu and Restrepo model the introduction of new tasks. New tasks create new employment opportunities for labor, and I return to a discussion of these new tasks in the next section.

It is also useful to allow for labor market imperfections: for example, modeled as a wage floor, $\underline{w}$, below which equilibrium wages cannot fall. Without this wage floor, the labor market would be competitive and all workers who want to work will find employment, although—depending on which tasks are automated—all employed workers may end up specializing in a narrow set of tasks. When there is such a wage floor and it is binding, some workers may become unemployed, and the extent of unemployment will depend on the extent of automation.

As I explain in the Appendix, automation creates opposing productivity and displacement effects. The productivity effect helps labor, because it reduces production costs and subsequently increases labor demand. The displacement effect harms labor because workers are displaced from the tasks they used to perform and may need to be reallocated to perform other tasks where their marginal product may be lower. Even when the productivity effect is substantial, automation always reduces the labor share—and thus increases inequality

between capital and labor. Worse, when the productivity effect is small, automation may reduce wages and employment. The productivity effect, in turn, will be small when automation happens in tasks where labor is already fairly productive and capital is not very productive. Hence, the technologies that are worse for labor are those that displace tasks through automation, but are themselves not very productive—what Acemoglu and Restrepo call "so-so technologies". When the wage floor is binding, the introduction of so-so automation technologies will reduce employment.

What are the welfare consequences of employment-reducing automation? In a perfectly competitive market—where workers are at the margin, indifferent between leisure and work, and when there are no other distributional concerns—an automation-induced decline in employment does not have first-order welfare consequences. In fact, it is straightforward to see that the competitive equilibrium would always maximize net output (defined as total production minus what is used up to produce capital). However, when there are labor market imperfections, such as those captured by the wage floor $\underline{w}$, then low-productivity automation reduces welfare—thus motivating the term "excessive automation". This can be seen with the following argument: because the productivity effect is approximately zero, gross output and profits do not increase (workers in marginal tasks are replaced by machines, but total costs have not changed). Yet, capital usage, which is costly, increases, and this reduces net output. At this point, reallocating marginal tasks away from capital towards labor—thus reducing automation—would increase net output.

Why is the equilibrium misaligned with social welfare maximization? The answer is related to the wage floor. Firms, when making their hiring and automation decisions, are responding to the market wage, $\underline{w}$, whereas a utilitarian social planner—seeking to maximize net surplus—should take into account the opportunity cost of labor, which is zero. This argument establishes that, when productivity effects are limited, there will be excessive automation. It also pinpoints one of the channels for this type of inefficiency: in economies with labor market imperfections, firms base their automation decisions on the higher wage rate, rather than the lower social opportunity cost of labor.

This argument also clarifies that automation is likely to be excessive and potentially welfare-reducing *especially when* it generates small or negligible productivity effects. If the productivity gains from automation had been large, net output would have increased, even if it displaced workers. Moreover, with a large productivity effect, there may not have been a decline in labor demand in the first place (see the Appendix).

The case for excessive automation is strengthened if there are other considerations favoring higher levels of employment. For example, if employed individuals generate positive external effects (on their families and communities or for democracy) relative to the unemployed, then the social planner may want to increase employment beyond the equilibrium level. Distributional concerns would also weigh in the same direction because, in general, automation helps firms and firm owners, while reducing the labor share. In addition, as shown in Autor et al. (2003) and Acemoglu and Restrepo (2021), automation boosts inequality across worker groups, creating another distributional cost.

What does this imply for AI? AI is a broad technological platform, and can be used for developing many different types of technologies. Automation, especially automation of various white-collar tasks and jobs with significant decision-making components, is one of these applications. If AI is used for automation, then the arguments outlined above would also imply that low-productivity AI may reduce welfare. Two key questions are, thus,

whether AI technologies are likely to be deployed as capital and algorithms substituting for labor in various tasks and whether this will generate small or large productivity gains. The evidence in Acemoglu, Autor et al. (2021) suggests that there has been a significant uptick in AI activities since 2016, and that much of this has been associated with task displacement. That paper also finds reduced hiring in establishments that adopt AI technologies, so the evidence is consistent with, though does not prove, the idea that new AI technologies may not sufficiently improve productivity. There are other reasons why productivity gains from AI may be small. Most importantly, AI technologies are being used in some tasks in which humans are quite good (e.g., natural language processing, facial recognition, problem-solving; see Acemoglu, 2021).

In summary, the general lessons from this section are:

1. Automation reduces the labor share and may also reduce the (average) wage and/or employment, and the latter outcome is more likely when productivity gains from automation are small.
2. When labor market imperfections create a wedge between the market wage and the social opportunity cost of labor, automation tends to be excessive and welfare-reducing, particularly when it also impacts employment negatively. This too is more likely to be the case when the productivity gains are small. The same considerations apply when there are non-market reasons for preferring high levels of employment (for example, because employed workers contribute more to their families, communities or society in general).
3. Because it increases the capital share and reduces the labor share, and because it boosts inequality among workers, automation may also be excessive from a welfare point of view due to distributional concerns
4. If AI is used predominantly for automation, it will have similar effects to other automation technologies, and depending on its productivity effects and relevant welfare criteria, it may have a negative impact on social welfare.

Direction of AI technology and its labor market consequences

Some implications of AI used for automation were previously discussed, but it was also noted that AI, as a broad technological platform, can be used for creating new tasks or increasing labor productivity as well. Acemoglu and Restrepo (2018, 2019) show that the introduction of new (labor-intensive) tasks also has two effects: a productivity effect (because it raises the productivity of labor) and a "reinstatement effect", rooted in the fact that it reinstates labor centrally into the production process.

The productivity effect is positive as usual (even if the exact sources of productivity gains from new tasks are different than those from automation). The reinstatement effect is also positive, because new tasks generate new employment opportunities for labor. As a result, new tasks always increase employment and/or wages. Moreover, the presence of the reinstatement effect implies that the wage bill increases proportionately more than the productivity gains, pushing up the labor share—the converse of the impact of automation.

Acemoglu and Restrepo (2019) have argued that the reason why wages grew robustly during the decades following World War II was that rapid automation in certain tasks went hand-in-hand with the introduction of sufficiently many new tasks, counterbalancing the labor market implications of automation.

Returning to the implications of AI, with the same argument, using AI for new tasks would be welfare-improving, especially when there are labor market imperfections or other considerations favoring higher levels of employment than in equilibrium. Furthermore, if AI boosts the creation of new tasks and improves human productivity, it could counterbalance some of the adverse effects of automation by other technologies (e.g., robotics or specialized software).

When AI can be used both for automation and for new task-creation, the pivotal question becomes how the balance between these two activities is determined; that is, the direction of technological change. Acemoglu and Restrepo (2018) provide a framework for the analysis of the equilibrium direction of technology. This framework clarifies how the direction of technology depends on factor prices and the labor share, and also emphasizes that the equilibrium and the optimal directions of technology will often differ. In particular, labor market imperfections not only promote too much automation—as we saw in the previous subsection—but also tend to generate an unbalanced composition of AI research between automation and new tasks (see also Klinova, 2024, in this *Handbook*).

There may also be reasons for distortions in the direction of technological change that go beyond the purely economic. In Acemoglu (2021) I emphasize that the direction of technology is partly shaped by the business models of leading firms and the aspirations of researchers. If these favor automation, the equilibrium may involve too much automation, even absent economic distortions. A related argument is that US corporations may have become too focused on cost-cutting, which might also encourage excessive automation. Acemoglu et al. (2020), on the other hand, show that the US tax code imposes a much higher marginal tax rate on labor than on equipment and software capital, thus favoring automation. This policy channel triggers both excessive adoption of automation technologies and disproportionate emphasis on new automation technologies in research and development.

AI as a technological platform could, in principle, boost efforts to create new tasks. Take education as an example. Current investments in this area are focused on using AI technologies for automated grading and the development of online learning tools to replace various tasks performed by teachers. Yet, AI can be deployed for creating new tasks and directly increasing teacher productivity as well. It can be used for adapting teaching material to the needs and attitudes of diverse students in real time, overcoming a major problem of classroom-based teaching—the fact that students have diverse strengths and weaknesses and find different parts of the curricula challenging (see the discussion in Acemoglu, 2024). Likewise, AI has many diverse applications in health that can personalize care and empower nurses and general practitioners to make more and better decisions in care delivery. These promising directions notwithstanding, AI may be more likely to aggravate excessive automation. The current trajectory in AI research is shaped by the visions of large tech companies, who are responsible for the majority of the spending on this technology. Many of these companies have business models centered on substituting algorithms for humans, which may make them focus excessively on using AI for automation. At the same time, many AI researchers focus on reaching "human parity" in narrow tasks as the main metric of success, which could create another powerful force towards automation, rather

than towards using this platform to create new tasks. Like other automation technologies, AI may also appeal to executives intent on cost-cutting. If there are additional tax breaks and favorable treatments for software in general and AI-related technologies, specifically, these may exacerbate the focus on automation (Acemoglu & Johnson, 2023).

Overall, even though there is no definitive evidence on this question, it is possible that the direction of technological change was already tilted too much towards automation even before AI, which may have exacerbated these trends. If so, one of the major, harmful effects of AI could be its labor market implications.

The general lessons from this discussion are, therefore:

1. AI could, in principle, be used to increase worker productivity and to expand the set of tasks in which humans have a comparative advantage, rather than focusing mainly on automation. If used in this way, AI may counterbalance some of the negative effects of automation on labor and may generate positive welfare effects and beneficial distributional outcomes.
2. But there is no guarantee that the composition of technological change, in general, and the balance of AI between automation and more worker-friendly task creation will be optimal. In fact, there are many possible distortions, some of them economic and some of them social, that encourage excessive automation using AI.

AI and human judgment

The arguments in the previous two sections are partly predicated on the notion that AI-based automation may not generate sweeping productivity gains, which could compensate for—or even undo—the displacement effects it creates. AI's most-enthusiastic boosters, on the other hand, believe that AI can bring huge productivity gains. One of the most powerful arguments for this outcome is that, as AI helps automate and improve (both cognitive and noncognitive) tasks that do not require human judgment and creativity, it will increase the demand for problem-solving tasks that require creativity and judgment and also free workers to focus on these tasks. Although seemingly plausible, I now suggest a potential reason why this expectation may be too optimistic and argue that, even when such reallocation takes place, AI-based automation may be excessive.

The main idea is simple (and a more-detailed treatment is provided in the Appendix). In many activities there are "economies of scope", meaning that individuals acquire knowledge from performing certain tasks that can then be used in other tasks. When there are such economies of scope, automation of a subset of tasks may reduce potential human productivity in the remaining tasks as well. Moreover, in this case, firms will often adopt AI in order to reduce their costs in the AI-suitable tasks—ignoring the effects that this will have on the productivity of workers in other tasks.

More generally, a finer division of labor and the reallocation of some tasks away from humans can be cost-reducing, but to the extent that human judgment improves when workers gain experience from dealing with a range of problems and recognizing different aspects of those problems, it may also come at a cost. When some aspects of the problem are delegated to AI, workers may lose their fluency with, and ability to understand, the

holistic aspects of relevant tasks, which can then reduce their productivity—even for the tasks in which they specialize. An extreme example of this phenomenon can be given from the learning of mathematical reasoning. Calculators are much better than humans in arithmetic. But if students stopped learning arithmetic altogether, delegating all such functions to calculators and software, their ability to engage in other type of mathematical and abstract reasoning may suffer. For this reason, most mathematical curricula still emphasize the learning of arithmetic. If delegating certain tasks to AI becomes similar to the hypothetical cessation of learning arithmetic, it may have significant costs.

In the presence of economies of scope, the market equilibrium may involve the automation of tasks that could reduce worker productivity and adversely affect the overall efficiency of the economy.

In summary, the general lessons from this short discussion are:

1. In addition to the costs of worker displacement, discussed earlier in this section, economies of scope across tasks may create additional costs from the use of AI technologies. In particular, the deployment of AI in various cognitive tasks that do not require a high degree of human judgment and creativity may enable workers to reallocate their time towards tasks that involve judgment and creativity. But if economies of scope are important for human productivity, AI may have additional costs.
2. The cost-minimization incentives of firms may encourage them to use AI technologies in ways that forgo economies of scope.

AI and Excessive Monitoring

Another use of AI-powered technologies is in worker monitoring, as exemplified by Amazon's warehouses and new monitoring systems for delivery workers. Here, too, employers' incentives to improve monitoring and collect information about their employees predates AI. But once again, AI may magnify their ability to do so. Some amount of monitoring by employers may be useful to improve worker incentives. However, I argue that increasing employer flexibility in this activity can also lead to inefficiently high levels of monitoring, for a very simple reason: at the margin, monitoring is a way of shifting rents away from workers towards employers, and thus is not socially valuable. But precisely because it shifts rents to firms and will often increase their profits, firms would have an incentive to engage in monitoring even when it is not socially efficient. In such situations, new technologies that extend employers' abilities to engage in monitoring may be socially harmful.

The economic force here is, potentially, quite general: AI, by enabling better control and use of information, provides one more tool to employers that can shift rents away from workers and towards themselves, leading to inefficiently high levels of rent-shifting activities.

Is this potential inefficiency relevant? Once again, there is little systematic evidence to suggest one way or another, but the fact that the US labor market has a "good jobs" problem, and that wages at the bottom of the distribution have fallen in real terms over the last several decades (Acemoglu, 2019; Acemoglu & Restrepo, 2021), suggests that it may be.

In summary, the general lessons from this analysis are:

1. AI technologies also create new opportunities for improved monitoring of workers. These technologies have first-order distributional consequences because they enable better monitoring and, thus, lower efficiency wages for workers.
2. Because, at the margin, the use of monitoring technologies transfers rents from workers to firms, monitoring will be excessive in equilibrium. By expanding monitoring opportunities, AI may thus create an additional social cost.

AI, Political Discourse, and Democracy

The 1990s witnessed a rapid strengthening of democracy around the world, in a pattern the political scientist Samuel Huntington (1991) called "The Third Wave". During this process, many Latin American, Asian, and African countries moved from nondemocratic regimes towards democracy and several others strengthened preexisting democratic institutions (Markoff, 1996). The last 15 years have witnessed a pronounced reversal of this process, however. Several countries have moved away from democracy, and perhaps even more surprisingly, Democratic institutions and norms have come under attack in numerous Western nations (Levitsky & Ziblatt, 2018; Snyder, 2017; Mishra, 2017; Applebaum, 2020) and the citizenry appears to be more polarized than in the recent past (Abramowitz, 2010; Judis, 2016). Some have pointed to social media and online communication as major contributing factors to these headwinds (e.g., Marantz, 2020). I now turn to a discussion of these issues. I focus on the effects of AI on communication, political participation, and democratic politics. As indicated at the beginning of this chapter, I will highlight several distinct, but related, mechanisms via which AI might degrade democratic discourse.

Echo chambers and polarization

Social media is often accused of promoting echo chambers in which individuals communicate with others who are like-minded, which might prevent them from being exposed to counter-attitudinal viewpoints, consequently exacerbating their biases. Cass Sunstein noted the potential dangers of echo chambers as early as 2001. He stated that encountering individuals with opposing opinions and arguments is "important partly to ensure against fragmentation and extremism, which are predictable outcomes of any situation in which like-minded people speak only with themselves", and emphasized that "many or most citizens should have a range of common experiences. Without shared experiences, a heterogeneous society will have a much more difficult time in addressing social problems." (Sunstein, 2001, p. 9). The recent documentary *The Social Dilemma* describes the situation as, "[t]he way to think about it is as 2.5 billion Truman Shows. Each person has their own reality with their own facts. Over time you have the false sense that everyone agrees with you because everyone in your news feed sounds just like you." AI is crucial to this new reality on social media. For example, the algorithms that sites like Facebook and Twitter use to decide what types of news and messages individuals will be exposed to are based on applying AI

techniques to the massive amount of data that these platforms collect (Alcott & Gentzkow, 2017; Guriev et al., 2022; Mosleh et al., 2021). Recent studies document that these algorithmic approaches are exacerbating the problem of misinformation on social media, for example by creating algorithmic "filter bubbles", whereby individuals are more frequently exposed to news that aligns with their priors and biases than to news that challenges those priors (Levy, 2021). As a result, Vosoughi et al. (2018) conclude that on social media there is a pattern of "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information".

I will discuss these issues, building on the approach in Acemoglu, Ozdaglar, et al. (2021), which suggests that social media platforms may endogenously create such echo chambers and moreso in the presence of extremist content. Imagine an online community in which each individual receives a news item, which may contain misinformation. This misinformation may take the form of completely fake news or it may involve some misrepresentation of facts. The primary, individual decision is whether to further share the news.

Let us assume that if an agent is found out to have shared misinformation, she incurs a cost. To avoid this, she may instead decide not to share the article or even call it out for containing misinformation.

Suppose that there are two sub-communities, one is left-wing and the other is right-wing. Suppose also that each member of one of these communities is more likely to be connected to and sharing items with other members of her community, but there are also some cross-community links. The extent of these cross-community links determines how much "homophily" there is. With high homophily, there are almost no cross-community links. With low homophily, cross-community links are more common. Finally, suppose that each news item contains some "message", which is relevant for political decisions and beliefs, and also a reliability score, which measures the likelihood that the news contains misinformation. How do individuals make their sharing decisions?

The answer depends on the type of message they receive and the extent of homophily in the online community. Take a right-wing agent receiving a left-wing message. Given her political beliefs, she is much more likely to think that this message contains misinformation than a right-wing message. Therefore, all else being equal, she is much more likely to dislike and not share it. This effect is further magnified when she expects to share it with other fellow right-wingers because they are also more likely to dislike a left-wing message and call it out for containing misinformation.

In contrast, consider this right-wing agent receiving a right-wing message. Now she is less suspicious of the message and is thus more likely to share it. Interestingly, homophily now has the opposite effect on sharing decisions: if the individual is in a homophilic network, she expects other right-wingers to share the news item, and as a result, she is more likely to share it herself, creating a viral spread.

These observations lead to the first, basic result of the setup: a news item is more likely to become viral when (1) it reaches individuals who have congruent beliefs, and when (2) this concordant news item is being shared with others with similar beliefs.

Now consider the problem of the platform on which this community is situated. Suppose that, via its choice of algorithms, the platform influences the degree of homophily (as in Acemoglu, Ozdaglar, et al., 2021). Suppose also that the platform's objective is to maximize engagement and hence, greater virality of the article is better for the platform. First suppose that most of the news items arriving from the outside have high reliability scores

and are also being distributed roughly equally between the two sub-communities (so that it is not only left-wing messages going to the left-wing community and likewise for the right-wing community). Then the platform's engagement-maximizing policy is likely to be to introduce between-community links, thus exposing each individual to news items from the other side, since these high-reliability items will not be immediately disliked and may even reach a large audience in the two communities.

In contrast, suppose we have a situation in which there are many low-reliability news items and, moreover, left-leaning news items from the outside go to the left-wing community and right-leaning news items go to the right-wing community. If the platform were interested in stopping misinformation, it could choose low homophily (so that right-wing articles go to the left-wing community as well, for example), and this would induce less sharing among the right-wing agents. It would also cause interruptions to virality when left-wingers discover the right-wing news item to contain misinformation (which is likely, given that the news item has low reliablity). These considerations imply that when news items have lower reliability, the platform would prefer to induce extreme homophily via its algorithms and propagate misinformation in order to maximize engagement.

This result, mimicking the main finding of Acemoglu, Ozdaglar, et al. (2021), is in some ways quite striking. It highlights that the platform has incentives to create endogenous echo chambers (or filter bubbles). Worse, this happens precisely when there are low-reliability news items, which are likely to contain misinformation and be distributed within the sub-communities in a "polarized fashion" (e.g., left-wing messages going to left-wing groups, etc.).

The role of AI technologies is, again, crucial. Without these technologies, the platform would not be able to determine users' biases and create relatively-homogeneous communities. It would also not be able to support the rapid dissemination of viral news items.

Broader social implications of these types of filter bubbles are easy to see as well. Suppose that individuals also update their beliefs on the basis of the news items. When left-wingers receive right-wing news items and the relevant news items have reasonable reliability scores, they will tend to moderate their beliefs, exactly as Sunstein (2001) envisaged in the quote above. In contrast, when left-wingers receive only left-wing news items in a filter bubble, this might lead to further polarization. On the basis of these considerations, Acemoglu, Ozdaglar, et al. (2021) suggest various interventions, including regulation and outside inspections, in order to discourage such filter bubbles and reduce polarization. I return to the issue of regulation later in this chapter.

In summary, this section leads to the following general lessons:

1. AI-powered social media presents a variety of new opportunities for connecting individuals and information sharing.
2. However, in the process, it may also distort individuals' willingnesses to share unreliable information. When social media creates echo chamber-like environments, in which individuals are much more likely to communicate with like-minded others, they become less careful and more likely to share news items that are consistent with their existing views and more willing to allow the circulation of misinformation.
3. Centrally, social media platforms that are focused on maximizing engagement have an incentive to create echo chambers (or "filter bubbles"), because interruptions of the circulation of news items with unreliable messages reduces engagement.

As a result, especially when there are more items with misinformation, platform incentives are diametrically opposed to social objectives.

Perils of online communication

The previous section argued that political discourse may be hampered because of the algorithmic policies of digital platforms. I now suggest that there might be reasons more-endemic to the nature of online communication that disadvantage political communication (see also Lanier, 2018; Tirole, 2021). The main argument is a version of the ideas developed in Acemoglu, Kleinberg et al. (2022), and here I provide a brief discussion (with more details in the Appendix).

The main ingredient for the effects discussed here: there is more mutual information about, and greater trust between, participants in a real-world social network than in an online social media network. This trust makes it more likely that the individual will be able to influence the beliefs of their acquaintances in the real-world social network via political discourse. When they expect that political communication is feasible, individuals may prefer it to gossip. In contrast, in online social media, most interactions are between people who have less knowledge about each other and less trust in the communicators. With the same reasoning, this will tilt the balance towards gossip and other non-political forms of communication. As a result, online communication may become inundated with gossip at the expense of useful political exchange.

The situation may be worse when we take into account that online interactions typically take the form of broadcast (rather than bilateral) communication. This would exacerbate the situation I have outlined here for two reasons. First, there may be many agents who may want to broadcast their views, competing for attention. Because broadcasting may be particularly attractive to extremists, they may be overrepresented among those engaged in such competition. This would endogenously increase the assessment of all the agents that political news is coming from extremists. Second, if there is heterogeneity in the utility of gossip across agents, those who value gossiping most may be the ones monopolizing the channels. In both cases, online communication becomes less effective as a way of sharing politically or socially relevant information. If political communication and news sharing in social networks are an important aspect of democratic politics, then the forces identified in the subsection also create new challenges for democracy, again rooted in the use of AI technologies.

The general lessons from this brief analysis can be summed up as follows:

1. Bilateral, offline communication, especially when the subject matter is political or social, relies on trust between parties. The trust that naturally exists in the context of in-person social networks may enable this type of communication.
2. When communication is taking place online and in multi-lateral settings, such as in modern social media platforms powered by AI technologies, this type of trust-based communication becomes harder. This may favor non-political messages, such as gossip, which then drive out political communication, with potentially deleterious effects for political discourse and democracy.

3. This potential barrier to online communication is exacerbated when there is competition for attention, which is encouraged by the broadcast or multi-lateral nature of online communication.

Big Brother effects

The previous two sections focused on how AI-powered social media and online platforms change the nature of communication, with potentially negative effects on the sharing of political information, which is the bedrock of democratic participation by citizens. In this section, I suggest that the other crucial pillar of democratic institutions, citizen protests, is likely to be hampered by AI technologies.

I have argued in Acemoglu and Robinson (2000, 2006) that protests, riots, and uprisings are critical for the emergence of democratic regimes (because the threats that they pose for power-holders in nondemocratic regimes induces democratization). This argument is relevant for democratization in currently authoritarian governments, such as China, Russia, and Iran. Many AI-based technologies, including facial recognition, tools for identifying online and offline dissent, and data collection software that underpins citizen-score schemes (such as the Chinese social-credit system) can significantly strengthen authoritarian governments and even make it near-impossible for opposition groups to organize. Such technologies can significantly increase the longevity of nondemocratic regimes.

The problem is not confined to these nondemocratic nations, however, and applies to the US and other countries with democratic institutions as well. A similar argument suggests that protests and civil disobedience are often critical for the functioning of democratic regimes. The civil rights movement in the US illustrates this vividly. Even though the US was democratic at the federal level, the Jim Crow South routinely violated the political, social, and economic rights of Black Americans. Democratic institutions in the North and, to the extent that that they existed, in the South did not create a natural impetus for these discriminations to cease. The turning point came with civil disobedience organized by various Black (and later multi-ethnic) civil society groups, such as the NAACP. Vitally, even federal politicians opposed to Jim Crow were not in favor of these protests initially, viewing them as disruptive political drawbacks, especially given that any federal action against Jim Crow practices would trigger backlash from Southern politicians (see the discussion in Acemoglu & Robinson, 2019). Without civil disobedience, protest, and other sources of bottom-up pressures, it is likely that reform of voting, civil, education, and discrimination laws in the American South would have been further delayed.

In the Appendix, I provide a simple model illustrating how AI technologies prevent political dissent and protest activities. When there is the threat of protest, civil disobedience, and other types of dissent, a nondemocratic elite (or even an elite within an imperfect democracy) will be tempted to make various compromises in order to prevent adverse reactions from civil society. But when AI enables governments and corporations to shut down dissent, it also discourages elite compromises. As a result, political decisions become more elite-biased and political distortions multiply.

Overall, the general lessons from this discussion are:

1. AI technologies can be used to improve governmental monitoring of, and preventative action against, protest activities.

2. Because the threat of protests has a disciplining role on nondemocratic governments, and even on some democratic governments, the shift of power away from civil society towards governments will weaken democracy and aggravate policy distortions.

Automation, social power and democracy

The previous subsection explained how the use of AI as a tool for controlling society and political dissent can have harmful effects on democratic politics. I argue that AI-powered automation can further weaken democracy and undermine social cohesion.

To develop this argument, let us first go back to the framework in Acemoglu and Robinson (2006), which emphasizes two types of conditions for the emergence and survival of democratic institutions. First, there must be enough discontent with nondemocratic regimes to generate sufficient demand for democracy. Second, democracy should not be too costly for the elite, who would otherwise prefer to use repression or other means to avoid sharing political power with the broader population. One aspect of this problem, which could be important, but was not analyzed in Acemoglu and Robinson (2006), is cooperation from workers. For example, if workers become disenchanted with the current regime or decide that they need to take action against current (economic or political) power-holders, they may choose not to cooperate with capital in their workplaces. When labor is essential for production, this withdrawal of cooperation could be very costly for capital. When capital owners are influential in the political system, they may then push for democratization or redistribution in order to placate labor.

The main argument in this section is that automation may make workers less indispensable in workplaces, and as such, it will tend to reduce their political power. Moreover, when technology enables the automation of most of the tasks previously performed by humans, workers could be made completely unnecessary in production, thereby obliterating their political power (see also Boix, 2024, in this *Handbook*).

I present a model formalizing this argument in the Appendix. The main lessons from these ideas and the model are:

1. Automation can generate an indirect negative impact on democracy and redistributive politics when ensuring cooperation from labor in workplaces is an important motivation for elites to make concessions to labor.
2. When automation brings only small productivity gains, it encourages the elite to reduce redistribution and make fewer democratic concessions. This will make policies less responsive to the majority's wishes and may further raise inequality.
3. Productivity benefits of automation may soften this effect, because an automation-driven increase in output raises the opportunity cost of losing labor's cooperation. There may also exist a sufficiently high level of automation such that, once we reach this level, labor becomes sufficiently irrelevant for production that the withdrawal of workers' cooperation ceases to be very costly. After this threshold, the elite may prefer to abandon democratic institutions and withhold any concessions—proceeding without the cooperation of workers—once again, with harmful effects on democracy, redistribution, and social cohesion.

Other potential costs

I will now briefly list a few other areas that may be important, but without providing as much detail or formal analysis.

The threat of AI and bargaining. Even when the actual labor market effects of AI discussed previously are not realized, the threat of adopting AI technologies may influence wages and inequality. Specifically, if there is bargaining and rent-sharing, employers may use the threat of AI-based automation as a way of increasing their bargaining power, which could have some of the same effects as actual automation and AI-powered monitoring.

Discrimination. Bias in AI has already received considerable attention. As AI gains greater importance in our social and economic lives, ensuring that popular algorithms are fair and unbiased has become vital. Existing studies show that simple AI algorithms can improve important public decisions, such as bail or sentencing, without increasing discrimination (e.g., Kleinberg et al., 2018). However, in most such applications, algorithms use data generated from biased agents and potentially discriminatory practices (e.g., Thompson, 2019). For example, significant portions of the police and the judiciary in the United States are generally thought to be biased against certain groups, such as Black Americans. In such situations, there is a danger that preexisting biases will become a fundamental part of AI algorithms. This may not only promote persistent bias and discrimination, but may in fact cement these biases more deeply in society via a process similar to the "signaling role of laws" (e.g., Posner, 2002). Indeed, if society starts trusting AI algorithms, their discriminatory choices may come to be accepted as more justifiable than when they were made by individual decision-makers.

Technocracy versus democracy. Advances in AI may create the temptation to delegate more and more public decisions—or even political decisions—to algorithms or, rather, to the technocrats designing and using these algorithms. Although reliance on these "experts" may be justifiable for certain decisions, an excessively relied-upon technocracy, without citizen input, may also start encroaching into political decisions—such as the extent of redistributive taxation or how much we should protect disadvantaged groups (Sandel, 2020). In this case, reliance on AI may further undermine democracy, amplifying the concerns highlighted in the previous section.

AI-powered weapons. AI technologies have already started to be incorporated into weapons and are advancing towards autonomous weapon systems. These new technologies will cause a host of ethical and social dilemmas, and should perhaps be regulated before prototypes are deployed or even fully developed. In addition to these ethical and social issues, AI-powered weapons may further strengthen governments against civil society, protesters, and even some opposition groups, adding to the concerns we discussed in the previous section.

The alignment problem. The potential downside of AI technologies that has received the most attention is the "alignment problem": the problem of ensuring that intelligent machines have objectives that are aligned with those of humanity. Although all of the harms of AI emphasized so far can be thought of as rooted in a misalignment between the current crop of AI technologies and broader societal objectives, many researchers and public intellectuals have been concerned with another aspect of misalignment:

machines reaching super-human capabilities and then, implicitly or explicitly, turning against humans (e.g., Bostrom, 2014; Russell 2019; Christian, 2019). My own view is that these concerns are somewhat overblown and often distract from shorter-term problems created by AI technologies (on which this chapter has focused), but naturally, they deserve careful consideration, monitoring, and preparation. The general misalignment problem is studied through the lens of economic externalities by Korinek and Balwit (2024) in this *Handbook*.

The international dimension: The current development of AI technologies is intertwined with international competition, especially between the US and China (Lee, 2018). A discussion of AI regulation has to take into account this international dimension. For example, it may not be sufficient for the US and Europe to start regulating the use of data or excessive automation when these issues remain almost completely unregulated in China. This suggests that the regulation of AI needs to have a fully-fledged international dimension and we may need to build new international organizations to coordinate and monitor deregulation of AI across the globe.

The Role of Technology Choice and Regulation

In the preceding discussions, I consider a number of theoretical arguments which suggest that the deployment of AI technologies may generate economic, political, and social costs. In this section, I highlight that, in all of these cases, the problems are not inherent to AI technologies, per se. Rather, the harms I have emphasized are caused by corporate and societal choices about how these technologies are deployed. Even though these costs are far-ranging, taking place in product markets, in labor markets, and in the realm of politics, they exhibit a number of commonalities, which I explore in this section. I also discuss some possible remedies. The general emphasis in this section will be on three main ideas:

1. The importance of choices, both on the use of existent AI technologies and on the direction of AI research. The costs I have modeled are not intrinsic to AI technologies, and are instead based upon how this new technological platform is currently being developed: to empower corporations and governments against citizens and workers.
2. The inadequacy of market solutions that mainly rely on increasing competition.
3. The need for regulation.

Let me start with the mechanisms discussed earlier. All three potential costs of AI turn on how AI technologies enable the use and control of data. In each one of these cases, a different way of distributing control rights over data would ameliorate or prevent most of the costs (Posner & Weyl, 2019). Let us start with data markets. The source of inefficiency in this case is the ability of platforms to find out information about others from the data that an individual shares. This then opens the way to potential misuses of data—for example, to

reduce the surplus of consumers or to violate their privacy in other ways. Effective regulation in this case could take one of two forms. First, as suggested in Acemoglu, Makhdoumi et al. (2022), it may be possible to strip away parts of the data of an individual in order to prevent or minimize information about others being leaked (although the details matter here, as simply anonymizing data is not sufficient). Second, more systematic regulations on how platforms can use the information they acquire would lessen the harmful effects working through privacy.

In contrast, increasing competition may not be sufficient, and not even useful, in this case. My analysis focused on a monopoly platform. Acemoglu, Makhdoumi et al. (2022) show that if there are two platforms competing to attract users, this may exacerbate the pernicious effects of data externalities.

Let me then turn to the problem that was next considered. The source of inefficiency in this case is the ability of platforms to use the data that individuals reveal about themselves in order to manipulate their weaknesses. If this misuse of data can be prevented or if consumers can be made more-aware of how data are being used, some of these costs could be prevented. Suppose, for example, that consumers are informed frequently that platforms know a lot about their preferences and may use this knowledge to (very effectively) market low-quality products to their users. There is no guarantee that such informational warnings will work for all consumers, but if they are displayed saliently and are specific (e.g., calibrated according to the group of individuals and relevant class of products), they may prevent some of the harms identified above. In this case, too, increasing competition would not be an effective solution. If two platforms are competing for consumers, but consumers continue to be semi-behavioral and fail to recognize the increasing platform capabilities, both platforms may exploit their abilities to offer products that have apparent short-term benefits but disproportionate long-term costs.

The issues are similar when we turn to economic forces, but now the implications of competition are more nuanced. In this case, effective regulation would prevent one of the firms from using the additional information it acquires to capture all of the consumer surplus. Methods like price controls and limits on price discrimination might achieve these ends, although clearly such regulation is far from straightforward. What happens if we can increase competition in this case? Greater competition that results from firm o also using AI methods to estimate its own past consumers' preferences and customize its services accordingly would not be necessarily useful. Now both firms become local monopolists, capturing all of the consumer surplus. However, if each firm can also acquire information about the other's customers and if collusion can be prevented, then they can be induced to compete fiercely, with better outcomes for consumers. This case thus highlights that, in some scenarios, fostering competition might have benefits—though only to a limited extent and only when some case-specific conditions are satisfied.

I would also like to emphasize the implications for the direction of AI research in this case. Suppose that AI researchers can devote their time to developing alternative applications of this broad technological platform. For example, some of them may be able to use AI to create tools that empower citizens or consumers, or develop new technologies for preserving privacy. All the same, if any one of the mechanisms related to the control and misuse of information are relevant, then this mechanism will also produce a powerful

demand for technologies that enable corporations to acquire and better-exploit this type of information. These effects are exacerbated when the ability of consumers to pay for alternative technologies is limited, relative to the resources in the hands of corporations. In such scenarios, the demand for "misuse of AI" will be transmitted to AI researchers, who may then respond by devoting their time to developing the AI technologies that corporations demand and by moving away from technologies that have greater social value or that empower consumers and citizens. This is a general point, which applies whether the harmful effects of AI are on the control of information, labor markets, or politics. It is for this reason that innovation, when unregulated, is unlikely to produce self-correcting dynamics. To the contrary, the demand for misuse of AI will typically distort the allocation of AI research across different applications, amplifying the social and economic costs of this research.

The same considerations apply even more-evidently in the models presented in the discussion of automation. If automation is excessive, increasing competition in the labor market would not be particularly useful. On the other hand, the demand for automation technologies from firms will tend to be strong, encouraging researchers to double down on using AI to develop automation technologies. Regulatory solutions are feasible, but may be more difficult to design and implement in this case. In theory, when automation is excessive and AI research is not being directed to creating new tasks, welfare-promoting regulation should discourage automation at the margin and encourage the creation of new labor-intensive tasks.

However, distinguishing marginal (low-productivity) and infra-marginal (higher-productivity) automation is difficult. Even more challenging is that regulators might have a hard time separating the AI that is used to create new tasks from the AI that is used to automate low-skill tasks and empower higher-skilled or managerial workers. But it is also possible to view these problems not as absolute barriers, but instead as measurement challenges. More research might shed light on how to distinguish different uses of AI in the labor market and might reveal new regulatory approaches for influencing the direction of AI research.

Finally, there are similar lessons from the models we discussed when considering politics and democracy, although there are also some new challenges related to the fact that the effects are now on political and democratic outcomes. For one, increasing competition is unlikely to be a very effective way of dealing with misaligned platform incentives. For example, if there are multiple social media platforms trying to maximize engagement, each may have incentives to create filter bubbles. Pro-competitive solutions may also be less effective when there are systemic issues, such as the widespread malfunctioning of democratic institutions.

The effects of these new technologies for democratic politics raises new conceptual issues as well. Most importantly, if the incorrect deployment of AI technologies is weakening democratic politics, developing after-the-fact regulatory solutions might become harder because democratic scrutiny of those who benefit from the distortionary use of AI technologies would also become more difficult. These considerations suggest a "precautionary regulatory principle"—an *ex ante* regulation slowing down the use of AI technologies, especially in domains where redressing the costs of AI become politically and socially more difficult after large-scale implementation (Acemoglu & Lensman, 2023). AI technologies impacting

political discourse and democratic politics may be prime candidates for the application of such a precautionary regulatory principle.*

Conclusion

In this chapter I have explored several potential economic, political, and social costs of the current path of AI technologies. I have suggested that if AI continues to be deployed along its current trajectory and remains unregulated, it could harm competition, consumer privacy, and consumer choice. It may excessively automate work, fuel inequality, inefficiently push down wages, and fail to improve productivity. It may also make political discourse increasingly distorted, cutting one of the lifelines of democracy. I have also mentioned several other potential social costs from the current path of AI research.

I should reemphasize that these potential harms are theoretical. There is much evidence indicating that not-all is well with the deployment of AI technologies and that the problems of increasing market power, disappearance of work, inequality, low wages, and meaningful challenges to democratic discourse and practice are all real—though we do not have sufficient evidence to be certain that AI has, thus far, been a serious contributor to these troubling trends.

Nevertheless, precisely because AI is a promising technological platform, which aims to transform every sector of the economy and every aspect of our social lives, it is imperative for us to study the possible downsides, especially given its current trajectory. It is in this spirit that I have discussed the potential costs of AI in this chapter.

My own belief is that several of these costs are real and we may see them multiply in the years to come. Empirical work exploring these issues is, therefore, greatly needed.

Beyond empirical work, we also need to understand the nature and the sources of these potential costs and how they can be prevented. It is for this reason that I have suggested various policy responses, in each case emphasizing that the costs are rooted in the way that corporations and governments choose to develop and use these technologies. Therefore, my conclusion is that the best way of preventing these costs is to regulate AI and redirect AI research away from these harmful endeavors and towards areas where AI can create new tasks that increase human productivity and new products that can empower workers and citizens. Of course, I realize that such a redirection is challenging. The regulation of AI is probably more difficult to design than the regulation of many other technologies—both because of its fast-changing, pervasive nature and because of the international dimension. We must also be careful because history is replete with instances of governments and powerful interest groups opposing new technologies with disastrous consequences for economic growth (e.g., Acemoglu & Robinson, 2012). Redirecting technological change does not mean withholding technological advances, though it clearly requires government policy and societal pressure to discourage net-detrimental investments and deployment at the margin, while encouraging the development of more-optimal AI technologies (Acemoglu, Autor et al., 2023). The centrality of these technologies to our future, and their potential harms, justify the need for these conversations.

* My belief in this precautionary regulatory principle is also the reason why I became one of the signatories of the call to halt the training of large language models for six months, circulated here https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Appendix

In this Appendix, I present sketches of several models aimed at clarifying the economic mechanisms discussed in the text.

Model for Too Much Data

Following Acemoglu, Makhdoumi et al. (2022), consider a community consisting of n agents/users interacting on a (monopoly) digital platform. Each agent i has a type denoted by x_i which is a realization of a random variable X_i, where the vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ has a joint normal distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ (and $\Sigma_{ii} = \sigma_i^2 > 0$ denoting the variance of individual i's type). Each user has some personal data, S_i, which are informative about her type. Personal data include both characteristics that are the individual's private information (unless she decides to share) and also data that she generates via her activity online and offline. Suppose $S_i = X_i + Z_i$ where Z_i is a normally-distributed independent random variable, $Z_i \sim \mathcal{N}(0,1)$.

Although Acemoglu, Makhdoumi et al. (2022) discuss various metrics, here I suppose that the relevant notion of information is mean square error (MSE). Then we can define *leaked information* about user i as the reduction in the mean square error of the best estimator of her type:

$$\mathcal{I}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}[(X_i - \hat{x}_i(\mathbf{S}_\mathbf{a}))^2],$$

where $\mathbf{S}$ is the vector of data the platform acquires, $\hat{x}_i(S)$ is the platform's estimate of the user's type given this information, and $a = (a_1, \ldots, a_n)$ is the data-sharing action profile of users (with $a_i = 0$ denoting no direct data-sharing and $a_i = 1$ corresponding to data-sharing). Then, the objective of the platform is to maximize:

$$\sum_i [\eta \mathcal{I}_i(a) - a_i p_i],$$

where p_i denotes payment ("price") to user i from the platform, which is made only when the individual in question shares her data directly (i.e., $a_i = 1$), and $\eta > 0$. The price could take the form of an actual payment for data shared or an indirect payment by the platform, for example, the provision of some free service or customization. This specification embeds the idea that the platform would like to acquire data in order to better forecast the type/behavior of users.

User i's objective is different. She may wish to protect her privacy and she obviously benefits from payments she receives. Thus her objective is to maximize:

$$\gamma \sum_{i' \neq i} \mathcal{I}_{i'}(\mathbf{a}) - v_i \mathcal{I}_i(\mathbf{a}) + a_i p_i.$$

The first term represents any positive direct externalities from the information of other users (for example, because this improves the quality of services that the individual receives and does not fully pay for) and thus $\gamma \geq 0$. The second term is the loss of privacy (capturing

both instrumental and intrinsic values of privacy). Hence $v_i \geq 0$ here denotes the value of privacy to user *i*. Finally, the last term denotes the payments she receives from the platform.

This framework allows data to create positive or negative total benefits. To illustrate this point, suppose that $v_i = v$. In that case, data creates aggregate (utilitarian) benefits, provided that $\eta + \gamma(n-1) > v$. In contrast, if $\eta + \gamma(n-1) < v$, the corporate control and use of data is socially wasteful (it creates more damage than good). But even in this case, as we will see, there may be data transactions and extensive use of data. In general, because v_i differs across agents, data about certain users may generate greater social benefits than the costs, while the revelation of data about others may be excessively costly.

In terms of market structure, the simplest option is to assume that the platform makes take-it-or-leave-it offers to users in order to acquire their data.

A key result, proved in Acemoglu, Makhdoumi et al. (2022), is that $\mathcal{I}_i(\mathbf{a})$ is monotone and submodular. The first property means that when an individual directly shares her data, this weakly increases the information that the platform has about all individuals, i.e., $\mathcal{I}_i(\mathbf{a}') \geq \mathcal{I}_i(\mathbf{a})$ whenever $\mathbf{a}' \geq \mathbf{a}$. Mathematically, the second implies that, for two action profiles $\mathbf{a}$ and $\mathbf{a}'$ with $\mathbf{a}'_{-i} \geq \mathbf{a}_{-i}$, we have:

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) \geq \mathcal{I}_i(a_i = 1, \mathbf{a}'_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}'_{-i}).$$

Economically, it means that the information transmitted by an individual who directly shares her data is less when there is more data-sharing by others.

I now illustrate the implications of this setup for data sharing and welfare using two simple examples. Consider first a platform with two users ($i = 1,2$) and suppose that $\gamma = 0$, $\eta = 1$, and $v_1 < 1$ so that the first user has a small value of privacy, but $v_2 > 1$, implying that, because of strong privacy concerns, it is socially beneficial to not share user 2's with the platform. Finally, suppose that the correlation coefficient between the data of the two users is $\rho > 0$. Because $v_1 < 1$, the platform will always purchase user 1's data. But this also implies that it will indirectly learn about user 2, given the correlation between the two users' data. If v_2 is sufficiently large, it is easy to see that it would be socially optimal to close-off data transactions and not allow user 1 to sell her data either. This is because, by selling her data, user 1 is indirectly revealing information about user 2, whose value of privacy is very large. This illustrates how data externalities lead to inefficiency. In fact, if v_2 is sufficiently large, the equilibrium—which always involves user 1 selling her data—can be arbitrarily inefficient.

Let us illustrate this possibility using the same example. Consider the edge case where the information of the two users is very highly correlated, i.e., $\rho \approx 1$. In this example, the platform will know almost everything relevant about user 2 from user 1's data. The important observation is that this data leakage about user 2 undermines the willingness of user 2 to protect her data. In fact, since user 1 is revealing almost everything about her, user 2 would be willing to sell her own data for a very low price. In this extreme case with $\rho \approx 1$, both the willingness of the platform to buy user 2's data and benefits user 2 receives from protecting her data are very small, and thus this price becomes approximately zero. But the disturbing part for data prices and the functioning of the market in this instance: once user 2 is selling her data, this also reveals the user 1's data almost perfectly, such that user 1 too can only charge a very low price for her data (despite the fact that she values her privacy as well, albeit less than user 2). As a result, the platform will be able to acquire both users' data at approximately zero price. This price, obviously, does not reflect users' value of privacy. They may both wish to protect their data and derive significant value from privacy. Nevertheless, the market will induce them

to sell it for close to zero price. Imagine once again that v_2 is sufficiently high. Then, despite this high value of privacy to one of the users, there will be a lot of data transactions, data prices will be near zero, and the equilibrium will be significantly (arbitrarily) inefficient. These consequences follow from submodularity.

As a second example, consider the case in which again $\gamma = 0$ and $\eta = 1$, but now there is no heterogeneity between the two users, so that $v_1 = v_2 = v > 1$. This configuration implies that neither user would like to sell their data (because their privacy is more important than the value of data to the platform). Nevertheless, it can be shown that as long as v is less than some threshold $\overline{v}$ (which is, itself, strictly greater than 1), there exists an equilibrium in which the platform buys the data of both users relatively cheaply. This is also a consequence of submodularity: when each user expects the other one to sell their data, they become less willing to protect their own data and more willing to sell it relatively cheaply. This locks both users into an equilibrium in which their data is less valuable than they would normally assume and, partly as a result, there is again too much data transaction.

In addition to leading to excessive data use and transactions, the externalities also shift the distribution of surplus in favor of the platform. To see this, suppose $v_1 = v_2 = v \leq 1$ and $\rho \approx 1$, so that it is now socially optimal for data to be used by the platform. It is straightforward to verify that, in equilibrium, data prices will again be equal to zero and, thus, all of the benefits from the use of data will be captured by the platform.

Model for Data and Unfair Competition

Suppose that, as in the basic Hotelling model, consumers are located uniformly across a line of length 1 and incur a cost—similar to a transport cost—when they purchase a product further away from their bliss point, represented by their location (see, e.g., Tirole, 1989). I assume that the utility of consumer i with location (or bliss point) i can be written as:

$$\alpha - \beta(x_i^f - i)^2 - p_i^f,$$

where $x_i^f \in [0,1]$ is the product of firm $f \in \{0,1\}$ and p_i^f is its (potentially) customized price for this consumer. Throughout, we normalize the cost of production to zero for both firms (regardless of whether they produce a standardized or customized product).

Let us interpret the two firms as two different websites, which consumers visit in order to purchase the good in question. Before AI, firms cannot observe the type of consumer and I assume that they cannot offer several products to a consumer that visits their websites. Thus, they will have to offer standardized products. This description implies that, in terms of timing, they first choose their product, and then, after observing each other's product choice, they set prices. Because each firm is offering a standard product and cannot observe consumer type, it will also set the same price for all consumers. This makes the pre-AI game identical to a two-stage Hotelling model, in which firms first choose their product type (equivalent to their location) and then compete on prices. Throughout, I assume that:

$$5\beta < 4\alpha, \tag{1}$$

which is sufficient to ensure that the market is covered and that the firms will not act as local monopolies. As usual, I focus on subgame perfect equilibria, but with a slight abuse of terminology, I refer to these as "equilibria".

It is straightforward to see that the unique equilibrium in this model, as in the baseline Hotelling model with quadratic transport costs, is maximal product differentiation (Tirole, 1989). In this setting, the two firms will offer products at the two ends of the line ($x^0 = 0$ and $x^1 = 1$) and set equilibrium prices given by $p^0 = p^1 = \beta$, sharing the market equally. For future reference, I also note that, in this equilibrium, total firm profits are equal to $\Pi^{\text{pre-AI}} = \pi^0 + \pi^1 = \beta$, consumer surplus is:

$$\begin{aligned} CS^{\text{pre-AI}} &= \alpha - 2\beta \int_0^{1/2} x^2 dx - \beta \\ &= \alpha - \frac{13}{12}\beta, \end{aligned}$$

where the first line of this expression uses the symmetry between the firms and consumers on the two sides of $1/2$.

After advances in AI, one of the firms (say firm 1) can use data from its previous customers (those with $i \geq 1/2$) to predict their type and customize their products and prices.[2] In particular, I assume that, in this post-AI environment, firm 1 can observe the type of any consumer $i \geq 1/2$ that visits its website and offers a customized bundle (x_i^1, p_i^1) to this consumer. For simplicity, let us assume that firm 0 cannot do so and also that firm 1 cannot simultaneously offer customized and standardized products. Now in equilibrium, firm 1 will offer each consumer with $i \geq 1/2$ a customized product $x_i^1 = i$. It will also charge higher prices. The exact form of the equilibrium depends on firm 0's product choice, which, given its inability to use the new AI technology, cannot be customized. It is straightforward to see that firm 0 will also change its product, because it no longer needs as much product differentiation (since firm 1 will be charging higher prices). The unique post-AI equilibrium is one in which firm 0 changes its standardized product to $x^0 = 1/4$. It then sets a price that makes the consumers that are farthest away from it indifferent between buying its product and not doing so, i.e.,[3]

$$p^0 = \alpha - \frac{\beta}{16}.$$

It is also straightforward to see that it is optimal for firm 1 to set:

$$p_i^1 = \alpha \text{ for all } i \geq \frac{1}{2},$$

thus capturing all of the consumer surplus from the consumers about whom it has data.[4] In this equilibrium, we have $\Pi^{\text{post-AI}} = \alpha - \frac{\beta}{32} > \Pi^{\text{pre-AI}}$ (which is guaranteed by (1)), while consumer surplus is now:

$$\begin{aligned} CS^{\text{post-AI}} &= \frac{\alpha}{2} - 2\beta \int_0^{1/4} x^2 dx - \frac{1}{2}\left(\alpha - \frac{\beta}{16}\right) \\ &= \frac{1}{48}\beta. \end{aligned}$$

As a consequence, consumer surplus is much lower in this case. This can be seen most-clearly by considering the limit, where $\beta \to 0$, in which case the pre-AI consumer surplus is maximal (approaching α), while the post-AI consumer surplus becomes minimal (approaching zero). The negative impact of AI technologies on consumer surplus has two

interrelated causes. First, firm 1 now uses its better prediction power to capture all the surplus from the consumers, even though it is, in principle, offering a better product and could have increased consumer welfare. Second, given firm 1's more-aggressive pricing, firm 0 is also able to capture more profits, reducing even the surplus of consumers whose data is not being used.

It is worth noting that, in the present model, there is no intensive margin of consumer choice and the market is covered (under (1)). As a result, AI does not affect quantity purchased, and even when it reduces consumer welfare, it increases utilitarian welfare—in particular, greater customization reduces "transport costs". The logic of the model highlights that this need not be the case when there is a quantity/intensive margin, because higher markups may inefficiently reduce quantity purchased. We will see, there are other reasons for inefficiency in similar environments.

Model for Behavioral Manipulation

I now present a model that shows how behavioral manipulation can take place, thanks to AI technologies and data about consumers. This model draws on Acemoglu, Makhdoumi et al. (2023), but instead of the continuous-time learning model developed in that paper, here I use a much simpler setting with two periods.

Consider the following dynamic setting with two periods ($t = 0,1$) and no discounting. Consumers have a choice between two products, x^1 and x^2, in both periods. They are initially uncertain about which one will yield higher utility. Suppose in particular that the true utility that a consumer gets from the product is either H or $L = 0$. The prior belief of individual i is that these two products will yield high utility for them is, respectively, q_i^1 and q_i^2.

Both products are produced and offered by a digital platform, which again has zero cost of production and can offer personalized prices. To start with, the platform and the consumer have symmetric information, and thus the platform knows and shares the consumer's prior beliefs. Once an individual consumes one of the two products, she obtains an additional piece of information about her utility from the product. I assume, in particular, that if the true quality is H, the consumer receives a positive signal, denoted by σ^H (with probability 1). However, if the true quality is L, the product might still have deceptively high instantaneous utility (but long-term costs). Thus, with probability λ, the consumer will receive the high signal (or will receive the low signal, σ^L, with complementary probability). The most relevant interpretation of this "false-positive" signal is that there are certain types of products that (predictably) appear more attractive to consumers—for example, because of their tempting short-term benefits or because of hidden negative attributes.

Let us assume that the platform perfectly observes the consumer's experience with the product that she has consumed, and can change its pricing and product offering in the next period. The game ends at the end of the second period.

The pre-AI equilibrium takes a simple form. The platform will offer whichever product has higher q_i for consumer i, say product j, and will set the price:

$$p_{i,0}^j = q_i^j H,$$

capturing the full surplus. If the signal after consumption is σ^L, then in the next period, it will offer the other product, $\sim j$, charging the lower price:

$$p^{\sim j}{}_{i,1} = q^{\sim j}{}_i H,$$

once again capturing the full surplus. If, on the other hand, the signal is σ^H, then, in the second period, the same product will be offered, but now there will be a higher price. I assume that, in the pre-AI environment, consumers have sufficient experience with such products and the signals they generate that they can correctly anticipate the likelihood of a high-quality product given a positive signal. As a result, the price following a positive signal will not increase all the way to H. Rather, it will be given by the expected value of the product's quality conditional on a positive signal. A simple use of Bayesian updating gives this price as:

$$\begin{aligned} p_{i,0}^j &= \frac{q_i^j}{q_i^j + (1-q_i^j)\lambda} H \\ &= \Xi_i^j H, \end{aligned}$$

which again captures the full surplus from the consumer and also defines the expression Ξ_i^j, which is convenient for the remainder of this section. (There is no option-value term in prices, because the full surplus is being captured by the platform.)

The deployment of AI technologies, once again, improves the platform's ability to predict consumer preferences and behavior because it has access to the data from many similar consumers and their experiences with similar products. As pointed out above, I assume that this goes beyond what the consumer herself knows. In particular, I suppose that the platform can now forecast whether the consumer will receive the high signal from a truly low-quality product. More generally, this captures the ability of the platform to predict whether the individual will engage in an impulse purchase or make other choices with apparent short-term benefits and long-term costs.

Post-AI, therefore, the relevant state for consumer i at time $t = 0$ becomes $\left(\left\{q_i^j, \xi_i^j\right\}_{j=1,2}\right)$, where $\xi_i^j = 1$ designates the event that product j will generate a false-positive signal—which means that, in reality, it is low-quality for the consumer, but still the signal σ^H will be realized if the consumer purchases it. Critically, the platform observes ξ_i^j, but the consumer does not. Following Acemoglu, Makhdoumi et al. (2023), I assume that consumers are "semi-behavioral" and do not fully take into account that, in the post-AI world, the platform actually knows ξ_i^j. This captures the more-general economic force mentioned above: in the pre-AI, business-as-usual world, consumers may have learned from their repeated experiences and purchases, accurately estimating the relevant probabilities. The post-AI world is new and it is less plausible to expect that the consumers will immediately understand the superior information that the platform has acquired. Note also that although it can forecast ξ_i^j, the platform cannot observe consumer preferences perfectly, and thus, even when $\xi_i^j = 0$, it does not know whether the product itself is high or low-quality.

What does equilibrium look like in the post-AI world? The key observation is that, while before AI the platform's prediction was aligned with the prior of the household, this is no longer the case in the post-AI world. In particular, suppose that we have $q_i^1 > q_i^2$, but $\xi_i^1 = 0$, while $\xi_i^2 = 1$. Then the platform may prefer to offer the second product. To understand this choice, let us compute the profits from consumer i when the platform is using these two strategies. When it offers product 1, its total profits are:

$$\pi_i^1 = \left[q_i^1 + q_i^1 \Xi_i^1 + (1-q_i^1) q_i^2\right] H.$$

This expression follows by noting that the platform is, at first, offering product 1 and charging q_i^1. Because $\xi_i^1 = 0$, the consumer will receive a positive signal only if the product is truly high quality, which happens with probability q_i^1. However, as indicated by the above discussion, in this case, the consumer does not know whether this is a false-positive or a truly high-quality product, and thus her valuation will be $\Xi_i^1 H$, which explains the second term. Finally, if she receives a negative signal (probability $1 - q_i^1$), in the second period, the platform will offer product 2, charging q_i^2.

On the other hand, when it initially offers product 2, the platform's profits are:

$$\pi_i^2 = (q_i^2 + \Xi_i^2)H,$$

because in this case there will be a positive signal for sure.

It is straightforward to see that offering the second good is more profitable for the platform when:

$$\frac{q_i^2}{q_i^2 + (1 - q_i^2)\lambda} > q_i^1(1 - q_i^2) + \frac{(q_i^1)^2}{q_i^1 + (1 - q_i^1)\lambda}. \tag{2}$$

Condition (2) is always satisfied whenever q_i^2 is sufficiently close to q_i^1. Intuitively, the platform is willing to sacrifice a little bit of revenue in the first period for the certainty of getting the consumer to experience a good that it knows she will like—even though this is not a truly high-quality good.

What about consumer welfare? Perhaps paradoxically, in the first case, the consumer actually has a positive welfare. This is because, in this case, we have a high-quality product (and the platform indirectly recognizes this following the realization of signal σ^H, because it knows that $\xi_i^1 = 0$) and hence the positive signal can come only from a truly high-quality good. It is then straightforward to compute the user's welfare as:

$$\begin{aligned} U_i^1 &= q_i^1 \left(1 - \frac{q_i^1}{q_i^1 + (1 - q_i^1)\lambda}\right) H \\ &= \frac{q_i^1(1 - q_i^1)\lambda}{q_i^1 + (1 - q_i^1)\lambda} H > 0. \end{aligned}$$

This positive surplus may appear as the good side of our behavioral assumption. But the platform's second strategy shows the dark side. With this strategy, the consumer will overpay in the second period (because, given $\xi_i^2 = 1$, the product is, in reality, low-quality). Hence her utility is:

$$U_i^2 = -\frac{q_i^2}{q_i^2 + (1 - q_i^2)\lambda} H < 0.$$

Therefore, the ability of the platform to predict the consumer's preferences and vulnerabilities leads to a situation in which the platform can increase its profits by marketing low-quality products that are likely to appeal to the consumer in the short run.

Model for Labor Market Consequences

Suppose there is a single good in the economy, Y, whose production requires the combination of a measure 1 of tasks:

$$Y = \left(\int_{N-1}^{N} Y(z)^{\frac{\sigma-1}{\sigma}} dz \right)^{\frac{\sigma}{\sigma-1}}, \tag{3}$$

where $Y(z)$ denotes the output of task z and $\sigma \geq 0$ is the elasticity of substitution between tasks. The key economic decision is the allocation of tasks to factors. Let me focus on just two factors, capital and labor, and suppose, as in Acemoglu and Restrepo (2018, 2019), that each factor has task-specific productivities, determining its comparative advantage, and only tasks $z \leq I$ can be automated, given the current level of automation technology. This implies:

$$Y(z) = \begin{cases} A^L \gamma^L(z) l(z) + A^K \gamma^K(z) k(z) & \text{if } z \in [N-1, I] \\ A^L \gamma^L(z) l(z) & \text{if } z \in (I, N]. \end{cases}$$

Here $l(z)$ and $k(z)$ denote the total labor and capital allocated to producing task z. The state of technology is captured by the following: factor-augmenting terms, A^L and A^K,which increase the productivity of the relevant factor uniformly in all tasks; task-specific productivities, $\gamma^L(z)$ and $\gamma^K(z)$, which increase the productivity of a factor in a specific task; the threshold for tasks that are feasible to automate, I; and the measure of new tasks, N. Let us assume that $\gamma^L(z) / \gamma^K(z)$ is increasing in z, so that labor has a *comparative advantage* in higher-indexed tasks. Suppose that capital is produced from the final good, with marginal cost R, which also gives its rental rate. Labor is inelastically supplied, with total supply given by L, and the equilibrium wage is denoted by w.

Acemoglu and Restrepo (2018, 2019) characterize the competitive equilibrium in this economy. Here I allow both competitive and rigid labor markets, by assuming that the wage cannot fall below some level $\underline{w}$. In this case, the equilibrium wage can written as:

$$w = \max\{\underline{w}, \text{MPL}(L)\},$$

where MPL (L) is the marginal product of labor when there is full employment at L. The wage floor may be a consequence of regulations, such as minimum wages and union-imposed minima, or may result from other labor market imperfections, such as efficiency wage considerations.

Let us first focus on how the marginal product of labor changes (without any wage floor). Following Acemoglu and Restrepo (2018), this is given by:

$$\begin{aligned} \frac{\partial \ln \text{MPL}(L)}{\partial I} = & \frac{\partial \ln Y(L,K)}{\partial I} && \text{(Productivity effect)} \\ & + \frac{1}{\sigma} \frac{1 - s^L}{1 - \Gamma(N,I)} \frac{\partial \ln \Gamma(N,I)}{\partial I} && \text{(Displacement effect)} \end{aligned} \tag{4}$$

where s^L denotes the labor share and $\Gamma(N,I) = \dfrac{\int_I^N \gamma^L(z)^{\sigma-1} dz}{\int_{N-1}^I \gamma^K(z)^{\sigma-1} dz + \int_I^N \gamma^L(z)^{\sigma-1} dz}$ is a measure of labor's *task content of production* (capturing what fraction of tasks are assigned to labor). In

the special cases where $\sigma = 1$ or where $\gamma^K(z) = \gamma^L(z)$, we have $\Gamma(N,I) = N - I$, but more generally, $\Gamma(N,I)$ is always increasing in N and decreasing in I. The first line of (4) represents the *productivity effect*, which is driven by the fact that automation reduces costs and, thus, increases productivity—by an amount equivalent to the cost difference between producing the marginal tasks by labor vs. capital:

$$\frac{\partial \ln Y(L,K)}{\partial I} = \frac{1}{\sigma - 1}\left[\left(\frac{R}{A^K \gamma^K(I)}\right)^{1-\sigma} - \left(\frac{\text{MPL}(L)}{A^L \gamma^L(I)}\right)^{1-\sigma}\right].$$

The second line of (4) represents the *displacement effect* created by automation: as tasks are allocated away from labor and towards capital, the marginal product of labor declines. This displacement effect, which reduces the range of tasks employing workers, is always negative.

When we are at full employment, (4) gives the impact of automation on wages. When, instead, the wage floor at $\underline{w}$ is binding, then the same effects now impact employment. The only differences are that on the left-hand side of (4), we now have the proportional change in employment, and on the right-hand side, $\text{MPL}(L)/A^L \gamma^L(I)$ is replaced by $\underline{w}/A^L \gamma^L(I)$.

Let us first consider full employment. What happens to the labor market equilibrium following additional automation? Equation (4) first shows that the labor share will always decline—because of the displacement effect, the wage will increase less than proportionately with productivity. Equally importantly, the wage level may fall as well. This is because the displacement effect can be larger than the productivity effect. In particular, when the productivity effect is small—for example, in the edge case where $\underline{w}/A^L \gamma^L(I) \approx R/A^K \gamma^K(I)$—there is no productivity effect, and the equilibrium wage will necessarily decline. When there is more than one type of labor, the same argument also implies that, though the wage of some groups may increase, the average wage may still fall (see Acemoglu & Restrepo, 2021).

This framework further clarifies why automation could reduce employment. Suppose the wage floor $\underline{w}$ is binding and, again, take the edge case where $\underline{w}/A^L \gamma^L(I) \approx R/A^K \gamma^K(I)$ so that the productivity effect is approximately zero. Then, automation necessarily reduces employment. By continuity, the same happens when the productivity effect is positive but not too large—the case that Acemoglu and Restrepo (2019) refer to as "so-so technologies", because they are good enough to be adopted, but not so good as to have a meaningful impact on productivity.

Finally, let us consider the effects of the introduction of new (labor-intensive) tasks, captured by an increase in N. As in Acemoglu and Restrepo (2018, 2019), we now have:

$$\begin{aligned}\frac{\partial \ln \text{MPL}(L)}{\partial N} = {} & \frac{\partial \ln Y(L,K)}{\partial N} && \text{(Productivity effect)} \\ & + \frac{1}{\sigma}\frac{1 - s^L}{1 - \Gamma(N,I)}\frac{\partial \ln \Gamma(N,I)}{\partial N}. && \text{(Reinstatement effect)}\end{aligned} \tag{5}$$

The productivity effect is positive as before and, in fact, introducing new tasks may increase productivity much more than automation. In addition, the reinstatement effect is also positive because, by creating new employment opportunities, it raises the labor share and increases employment and wages. When employment is suboptimally low because of labor market imperfections, new tasks and the reinstatement effect will be welfare-increasing.

Model for AI and Human Judgment

Suppose that there are two tasks to be performed, 1 and 2. Overall output in the economy is given by:

$$Y = \min\{y^1, y^2\},$$

where y^i is the output of task i, and the Leontief production function imposes that these tasks are strongly complementary.

Before AI, both tasks have to be performed by humans. Suppose that there is a measure 1 of humans, each with 2 units of time. Suppose also that, for reasons I will explain below, we start with humans allocating half of their time to task 1 and the other half to task 2. In this case, they have equal productivity in both tasks, which I normalize to 1. As a result, before AI, the economy produces a total of one unit of the final good. We can think of each worker as a "yeoman-producer", consuming his or her production. Equivalently, we can think of this economy as consisting of firms hiring workers in a competitive labor market. In this case, the per-hour wage of each worker in each task will be $1/2$, ensuring that the entire output is paid to workers.

Now imagine that there are advances in AI algorithms that produce the first task at per-unit cost $c < 1/2$. This cost is paid in terms of the final good, and the fact that it is less than the equilibrium wage before AI implies that these algorithms are cost-saving and will be adopted. If this were the end of the story, AI would improve net output because workers would be reallocated from task 1 to task 2, enabling the economy to increase its total output.

However, suppose that there are also economies of scope: individuals learn from performing both tasks at the same time (and that is why the pre-AI allocation involved each worker devoting half of their time to each task). Suppose, in particular, that if a worker does not learn from task 1, his or her productivity in task 2 declines to $1-\beta$. The post-AI allocation will involve all workers working in task 2 and whatever their total production is in this task, the economy will also produce exactly the same amount of task 1, using AI algorithms. As a result, net output in this economy will be:

$$2(1-\beta) - \text{spending on AI} = 2(1-\beta)(1-c).$$

It can be verified that in the special case where there are no economies of scope ($\beta = 0$), $c < 1/2$ is sufficient for net output to increase—in particular, from 1 to $2(1-c) > 1$. However, as soon as $\beta > 0$, this is no longer guaranteed. For example, when $c \approx 1/2$, even small economies of scope imply that the use of AI would reduce net output.

Would the market adopt AI when there are such economies of scope? The answer depends on the exact market structure. Because the adoption of AI technologies is associated with a finer division of labor, there is no guarantee that firms will internalize the economies of scope. For example, in the pre-AI equilibrium, the cost of one unit of task 1 is $1/2$. So, a new firm can enter with the AI technology and make profits in this equilibrium. The entry of these firms would then create a pecuniary externality, discouraging other workers from working on task 1. In particular, even though there are economies-of-scope benefits from performing task 1, workers may not be allocated to task 1 because the price of this task is now lower due to the use of AI technologies.[5] This type of entry could then destroy the pre-AI equilibrium and drive the economy to the post-AI equilibrium characterized above, even when it is inefficient because β is large.

Model for AI and Excessive Monitoring

I now develop the main ideas discussed in the text using a model based on Acemoglu and Newman (2002). Consider a one-period economy consisting of a continuum of measure N of workers and a continuum of measure 1 of firm owners, each with a production function $AF(L_i)$ where L_i denotes the number of workers employed by firm i who exert effort (the alternative, exerting zero effort, leads to zero productivity). Firms are large, so the output of an individual worker is not observable. Instead, employers can directly monitor effort in order to determine whether an employee is exerting effort and being productive.

Specifically, as in Shapiro and Stiglitz (1984), a worker exerting effort is never mistakenly identified as a shirker, and a shirking worker is caught with probability $q_i = q(m_i)$ where m_i is the extent of monitoring per worker by firm i, with cost Cm_iL_i, where $C > 0$. Suppose that q is increasing, concave and differentiable with $q(0) = 0$ and $q(m) < 1$ for all m. Suppose also that, because of the limited liability constraints, workers cannot be paid a negative wage and will simply receive a zero wage. This implies a simple incentive compatibility constraint for workers,

$$w_i - e \geq (1 - q_i)w_i,$$

where e denotes the cost of effort. Rearranging this equation we obtain:

$$w_i \geq \frac{e}{q(m_i)}. \tag{6}$$

In addition, the firm has to respect the participation constraint,

$$w_i - e \geq \underline{u}, \tag{7}$$

where $\underline{u}$ is the worker's *ex ante* reservation utility, given by what he could receive from another firm in this market.

The firm maximizes its profits, given by $\max_{w_i, L_i, q_i} \Pi = AF(L_i) - w_iL_i - Cm_iL_i$, subject to these two constraints. As shown in Acemoglu and Newman (2002), the solution to this problem takes a simple form because the incentive compatibility constraint always binds—if it did not, the firm would reduce monitoring, increasing its profits. By contrast, the participation constraint (7) may or may not bind.

The main result from this framework relevant for my discussion, is that regardless of whether the participation constraint is binding, the equilibrium never maximizes utilitarian social welfare (total surplus), given by $Y = AF(L) - CmL - eL$, because there is always too much monitoring. The intuition is straightforward: at the margin, monitoring is used to transfer rents from workers to firms, and is thus used excessively. Mathematically, this can be seen in the following way: start from the equilibrium and consider a small decline in monitoring, coupled with a small increase in the wage, such that (6) remains binding. This will have only a second-order impact on the firm's profits because the firm was already maximizing them. But it will have a first-order benefit for workers, whose wages will increase. Thus, a utilitarian social planner would like to increase wages and reduce monitoring. This is true, *a fortiori*, if we care about income inequality (presuming, of course, that firm owners are richer than workers).

How does AI affect things in this equilibrium? Suppose that before AI, there was an upper bound on monitoring, so that $m \leq \bar{m}$, and AI lifts this constraint. If the equilibrium level of monitoring without this constraint, m^*, is above $\bar{m}$, then improvements in AI will lead to higher monitoring. If, in addition, $\bar{m}$ is not too low, then AI would reduce welfare; the qualifier that $\bar{m}$ should not be too low is included because, if it were very low, the initial equilibrium could be very inefficient.

Model for Perils of Online Communication

Let us first focus on bilateral communication and consider a pre-AI world in which large-scale social media is not possible. Suppose that, in such a world, individuals communicate bilaterally in a social network. For simplicity, we can imagine a social network that takes the form of a directed line, in which we start with individual 0, who can communicate with individual 1, and so on, all the way up to the end of the line, individual n. Suppose that each individual has a piece of gossip which they can share with their neighbor, which will give utility v^G. In addition, the individual may have some news relevant to the political/social beliefs of the community. If she decides to share this news item, which then alters the beliefs of her neighbor, she also receives utility from such persuasion, as I specify below (for simplicity, I am ignoring potential utility benefits from non-neighbors indirectly changing their beliefs as this information is shared further). Because there are "channel constraints" (for example, limited ability to communicate), the individual can share either political news or gossip, but not both.

Specifically, suppose that the state of the world is 0 or 1 (Left or Right), and all individuals start with prior $\mu_0 = \frac{1}{2}$ about the underlying state being $s = 0$. Suppose that individual 0 receives a piece of news that shows that the underlying state is in fact $s = 0$. If she shares this information with her neighbor (individual 1) and her neighbor believes it, then her neighbor's belief will also shift to $\mu = 1$. However, each individual is also concerned that some agents in society may have ulterior motives and try to convince them that the state is the opposite of the true state, s. Suppose the probability that each individual attaches to their neighbor being of this extremist type is q. Then, the posterior of individual 1 that the state is $s = 0$ after receiving this news will be:

$$\mu_1 = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}q} = \frac{1}{1+q} > \frac{1}{2}.$$

Finally, I assume that individuals receive additional utility from shifting their neighbor's beliefs towards the truth (or her own belief), so the overall utility of individual i is:

$$v^G x_i^G + v^N \left| \mu_{i+1} - \mu_i \right|,$$

where $x_i^G = 1$ denotes whether this individual gossips, μ_i is her belief, and μ_{i+1} is her neighbor's belief (given the line network).

Let us assume that $v^N > 2v^G$, such that, if an individual is convinced that her information will be believed and can thus shift her neighbor's belief from $1/2$ to her views, she prefers to share the political information rather than gossip. Therefore, there exists $\bar{q}$ such that if $q < \bar{q}$, the individual will share the political news. Let us suppose that in the real-world social network this inequality is satisfied.

In this scenario, individual 0 will start sharing the political news. This will convince individual 1, who will then attach a sufficiently high probability to the underlying state being $s = 0$.

With the same argument, she would also like to convince her neighbor to the right, individual 2. If q is sufficiently small, then individual 2, even when she is worried about the possibility that either individual 0 or individual 1 is an extremist, would still believe this information. In general, there will exist some $n(q)$, such that the political news will be communicated up to individual $n(q)$, and then after that there will be pure gossip on the network. For q sufficiently small, the entire network might share the political news.

Next suppose that we go to online communication through social media networks. This offers a larger network with less personal contact. As a result, it is plausible to assume that the probability that each agent attaches to the event that the person communicating with them is an extremist is now higher, say $q' > q$. If $q' > \overline{q}$, then in online communication, there will be no news exchange and all communication will be gossip. The further distortions that arise in the presence of multi-lateral (broadcast) communication are discussed in the text.

Model for Big Brother Effects

Here I briefly outline a simple model capturing some of these ideas. Consider a society consisting of $\lambda < 1/2$ elites and measure 1 of regular citizens. All citizens and all elites have the same economic preferences, but citizens are heterogeneous in terms of their cost of participating in protest activities, denoted by c_i for individual i. I assume that c is distributed uniformly over [0,1] in the population.

The political system is an imperfect democracy or an autocracy in which political choices are biased in favor of the preferences of the elite. In particular, suppose that there is a unique, one-dimensional policy, and the preferred policy choice of the citizens is 0, while the most-preferred policy of the elite is $p^E > 0$. Consider the following reduced-form political game. The elite decide a policy, p, and then protests take place. If some number, q, of the citizens protest and engage in civil disobedience, then there is probability $\pi(q)$ that the policy will switch from p to 0. With the complementary probability, the policy stays at p. This political structure ignores the influence of the citizens via democratic institutions, which is for simplicity. If this is incorporated (e.g., Acemoglu & Robinson, 2008) this would not affect the main message of the model presented here.

I also assume that the government can impose a punishment on those engaged in protests. Suppose, in particular, that the state has the capacity to detect at most a measure ψ of protesters. If the total amount of protest, q, is less than ψ, then all protesters are detected and can be punished. I assume that the punishment imposed on protesters is a constant, Γ, independent of the number of protesters.

The two key economic decisions are, therefore, policy choice by elites and protests by citizens. Let me first describe the utility of the citizens. Suppose that when the policy choice of elites is p and there are q protesters in total, individual i has the following utility as a function of her protest decision $x_i \in \{0,1\}$:

$$U_i^C(p,q,x_i) = \left[\left(v^C |p| - c_i \right) - \min\left\{ \frac{\psi}{q}, 1 \right\} \Gamma \right] x_i,$$

where, for centricity, I have ignored components of the utility that depend on policy choice but are independent of the individual's process decision. Intuitively, the utility from protesting is increasing in the distance between the actual policy and the bliss point of citizens, as captured by $v^C|p|$ (recall that their bliss point is at zero). In addition, the individual incurs the cost of

participating in protests, given by c_i. The second term in square brackets captures the expected punishment from protesting, taking into account that, when $q \leq \psi$, protesters will be punished with probability 1.

Clearly, there exists a threshold value $\overline{c}$, such that only individuals with $c_i \leq \overline{c}$ will participate, and thus $q = \overline{c}$, since the distribution of c is uniform between 0 and 1.

Let us next turn to the elite's utility. Suppose that this is given by:

$$\begin{aligned} U^E(p,q) &= -\mathbb{E}\left[\left|\hat{p} - p^E\right|\right] \\ &= -(1-\pi(q))v^E\left|\hat{p} - p^E\right| - \pi(q)v^E\left|p^E\right|, \end{aligned}$$

where $\hat{p}$ denotes the realized policy and the expectation is over the uncertainty concerning whether protests will force the elites to change policy. As usual, the subgame perfect equilibrium can be solved by backward induction. In the second stage, $\overline{c}$ is determined such that, given the policy choice of elites, p, we have:

$$\left(v^C|p| - \overline{c}(p)\right) - \min\left\{\frac{\psi}{\overline{c}(p)}, 1\right\}\Gamma = 0. \tag{8}$$

In general, there can be multiple equilibria in this stage, because this equation might have multiple solutions for $\overline{c}(p)$. Note, in particular, that its left-hand side may be non-monotonic. In what follows, I focus on the case in which it is monotonically decreasing, which ensures a unique equilibrium; if there were multiple equilibria, however, we could pick the one with the highest amount of protest, which will necessarily be one where the left-hand side is decreasing, yielding the same results. It is then straightforward to see that $\overline{c}(p)$ is increasing in p, meaning that a more pro-elite policy will induce more protest.

Now, turning to the elite's maximization, we first rewrite the elite's utility function, taking into account the reaction of the citizens to their policy choice:

$$U^E(p,\overline{c}(p)) = -(1-\pi(\overline{c}(p)))\left|p - p^E\right| - \pi(\overline{c}(p))\left|p^E\right|,$$

which simply substitutes $q = \overline{c}(p)$. We can, therefore, maximize elite utility by choosing the initial policy p. Taking into account that $p < p^E$, this maximization problem yields a standard first-order condition:

$$(1-\pi(\overline{c}(p))) - \pi'\left(\overline{c}(p)\right)\overline{c}'(p)(2p^E - p) = 0, \tag{9}$$

and under suitable assumptions, we can ensure that the second-order condition for maximization is satisfied. In this first-order condition, $\overline{c}'(p)$ is given from (8) and, under the assumption that $\psi < \overline{c}(p)$, it can be written as:

$$\overline{c}'(p) = \frac{v^C}{1 + \psi\Gamma / \overline{c}(p)^2}.$$

Next, consider the introduction of AI, modeled as an increase in ψ to some ψ'. From the previous expression, this increase will reduce $\overline{c}'(p)$ and, from (9), this reduction in $\overline{c}'(p)$ will increase p away from the citizenry's bliss point and towards the elite's preferences. Intuitively,

AI-induced government monitoring of protests weakens citizens' collective ability to force the elites to make concessions, so the elite respond to the deployment of the AI technology by withdrawing concessions. As a result, AI makes policies less-responsive to the citizens' wishes and to the extent that these policies impact the distribution of resources, it will also tend to favor the elite's economic interests and increase inequality.

Model for Automation, Social Power and Democracy

Let us return to the model for Labor Market Consequences and, for simplicity, suppose that $N=1$ and $\sigma=1$ in the production function (3) and that labor markets are competitive (i.e., there is no wage floor). This implies that the equilibrium level of production, as a function of capital and labor, can be written as:

$$Y(K,L)=K^{I}L^{1-I},$$

and, thus, with competitive labor markets, the labor share is $s^{L}=1-I$ and:

$$w=(1-I)\frac{Y(K,L)}{L}.$$

Note also that, in this case, the impact of automation on output can be written as:

$$\frac{\partial \ln Y(K,L)}{\partial I}=\ln K-\ln L.$$

Therefore, automation is output-increasing, when $\ln K>\ln L$, or $K>L$, which is also equivalent to the competitive rental rate of capital being less than the wage, ensuring that automation is cost-saving. Conversely, low-productivity ("so-so") automation now corresponds to the case in which $K\approx L$.

Consider a political system, as in Acemoglu and Robinson (2006), where all capital owners (capitalists) are elites and all workers are non-elites, with no within-group heterogeneity. To start with, let us consider a nondemocratic regime in which the capitalists hold power or, alternatively, a democratic regime in which they have disproportionate power. For brevity, I am going to ignore any threat of revolution or protests along the lines of the models considered in Acemoglu and Robinson (2006) and will also abstract from the considerations discussed in the previous subsection.

Suppose, in addition, that there is a lump-sum tax on capitalists, which can be redistributed to workers, and let us denote the per-worker transfer on the basis of this by τ. Suppose that the workers have an aspiration for a level of net income, given by w^{A}, and if $w+\tau<w^{A}$, they withdraw their cooperation, and as a result, the effective productivity of labor declines from 1 to $\delta<1$. In this reduced-form model, I interpret the transfer τ from the elite both as a measure of redistributive politics and also as a general concession to democratic politics—for example, allowing workers to have more voice or making less use of lobbying and other activities that distort democratic politics.

The key question is whether the elite will make the necessary transfers to convince workers to continue cooperating in workplaces. This boils down to the comparison of the following two options for the elite: redistribute via τ so that the aspirations of the workers are met, or

make due with lower labor productivity as a result of broken cooperation. Let us write the payoffs to the elite from these two strategies. Suppose that there is no within-elite heterogeneity, so it is sufficient to look at the overall income of capital. When the capitalists choose to meet the aspiration constraint of workers, their payoff is:

$$U_1^K = K^I L^{1-I} - \max\{(1-I)K^I L^{1-I}, w^A L\},$$

where the first term in the max operator applies when the market wage is already greater than w^A, while the second term is when they have to make transfers in order to bring workers to this level. Clearly, the relevant case for the discussion here is the latter, so I assume $(1-I)K^I L^{1-I} < w^A L$, and thus:

$$U_1^K = K^I L^{1-I} - w^A L.$$

The alternative is to not make the necessary transfer, in which case the elite simply receive the market return to capital when the productivity of labor has been reduced to δ, i.e.,

$$U_2^K = K^I (\delta L)^{1-I}.$$

What is the effect of automation on the comparison of these two strategies? Differentiating the difference in their payoffs, $U_1^K - U_2^K$, with respect to I yields:

$$\frac{\partial\left(U_1^K - U_2^K\right)}{\partial I} < 0 \text{ if and only if } \ln K - \ln L + \frac{\ln \delta}{1-\delta^{1-I}} < 0.$$

This expression has two immediate implications. First, if we have low-productivity automation, such that $\ln K \approx \ln L$, then, because $\ln \delta < 0$, automation always makes the second strategy of stopping redistribution and foregoing labor's cooperation more attractive. Intuitively, automation makes labor less-central for production and, thus, losing its cooperation becomes less costly for capital. The economic force going against this calculation is that when automation increases productivity (i.e., when $\ln K > \ln L$), the output loss due to lack of cooperation from labor becomes more costly. Second, however, we can also see that, for any K and L, there exists a threshold level I^* such that once $I > I^*$, the second strategy is preferred, even taking into account the productivity gains from automation. Therefore, automation makes cooperation from workers less important and, to the extent that securing this cooperation was an important part of the motivation for redistribution and democratic politics, automation may make the elites turn their backs on, or even become hostile to, democracy.

Notes

* I am grateful to many co-authors who have contributed to my thinking on these topics and on whose work I have heavily relied in this essay. They include David Autor, Jonathon Hazell, Simon Johnson, Jon Kleinberg, Anton Korinek, Ali Makhdoumi, Azarakhsh Malekian, Andrea Manera, Sendhil Mullainathan, Andrew Newman, Asu Ozdaglar, Pascual Restrepo and James Siderius. I am grateful to David Autor, Lauren Fahey, Austin Lentsch, Vincent Rollet, James Siderius and Glen Weyl for comments.

1. The field of "AI" today is dominated by the suite of current artificial intelligence technologies and approaches, mostly based on statistical pattern recognition, machine learning, and big data methods. The potential harms of AI I discuss in this paper are relevant for and motivated by these approaches. Nevertheless, I will also emphasize that "AI" should be thought of as a broad technological platform, precisely because the general aspiration to produce "machine intelligence" includes efforts to improve machines in order to complement humans, create new tasks and services, and generate novel communication and collaboration possibilities.
2. More generally, the fact that AI-intensive firms are using data from and customizing products to their existing customers introduces intertemporal linkages, which could create lock-in effects and rich-get-richer dynamics, as in the switching cost and dynamic oligopoly literatures, such as in Klemperer (1995) and Budd, Harris and Vickers (1993).
3. Firm 0 could offer a lower price and steal some customers from firm 1, but it can be verified that this would lead to lower profits.
4. If we had allowed this firm to also market a standardized product, it would additionally compete for consumers $i < 1/2$, about whom it has no data. Our assumption rules out this possibility.
5. There are some market structures and pricing schemes that may prevent the adoption of AI technologies when they are inefficient in this case. For example, if workers can take a very low or even negative wage in order to work in task 1 (so as to increase their productivity in task 2), this may outweigh the cost advantage of new firms that enter and specialize in using AI for task 1. The issues are similar to the ones that arise in the context of firm-sponsored general training, and as in that case, labor and credit market imperfections would typically preclude the possibility that workers fully pay for all of the benefits that they receive by taking wage cuts (see Acemoglu and Pischke, 1999).

References

Abramowitz, Alan I. (2010). *The disappearing center: Engaged citizens, polarization, and American democracy*. Yale University Press.

Acemoglu, Daron. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature 40*(1), 7–72.

Acemoglu, Daron. (2019). It's good jobs stupid. *Economics for Inclusive Prosperity Research Brief*. June. https://econfip.org/policy-briefs/its-good-jobs-stupid.

Acemoglu, Daron. (2021). AI's future doesn't have to be dystopian. *Boston Review*, May 20. https://www.bostonreview.net/forum/ais-future-doesnt-have-to-be-dystopian.

Acemoglu, Daron, & Autor, David H. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of labor economics, volume 4* (pp. 1043–1171). Elsevier.

Acemoglu, Daron, Autor, David, & Johnson, Simon. (2023). Can we have pro-worker AI? Choosing a path of machines in service of minds. *MIT Shaping the Future of Work, policy memo*. https://shapingwork.mit.edu/wp-content/uploads/2023/09/Pro-Worker-AI-Policy-Memo.pdf.

Acemoglu, Daron, & Johnson, Simon. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity*. PublicAffairs.

Acemoglu, Daron, & Lensman, Todd. (2023). Regulating transformative technologies. *NBER Working Paper No. 31461*.

Acemoglu, Daron, & Newman, Andrew F. (2002). The labor market and corporate structure. *European Economic Review 46*(10), 1733–1756.
Acemoglu, Daron, & Pischke, Jorn-Steffen. (1999). The structure of wages and investment in general training. *Journal of Political Economy 107*(3), 539–572.
Acemoglu, Daron, & Restrepo, Pascual. (2018). The race between man and machine: Implications of technology for growth, factor shares and employment. *American Economic Review 108*(6), 1488–1542.
Acemoglu, Daron, & Restrepo, Pascual. (2019). Automation and new tasks: How technology changes labor demand. *Journal of Economic Perspectives 33*(2), 3–30.
Acemoglu, Daron, & Restrepo, Pascual. (2021). Tasks, automation and the rise in US wage inequality. NBER Working Paper No. 28920.
Acemoglu, Daron, & Robinson, James A. (2000). Why did the west extend the franchise? Democracy, inequality and growth in historical perspective. *Quarterly Journal of Economics 115*(4), 1167–1199.
Acemoglu, Daron, & Robinson, James A. (2006). *Economic origins of dictatorship and democracy*. Cambridge University Press.
Acemoglu, Daron, & Robinson, James A. (2008). Persistence of power, elites and institutions. *American Economic Review 98*(1), 267–293.
Acemoglu, Daron, & Robinson, James A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Business.
Acemoglu, Daron, & Robinson, James A. (2019). *The narrow corridor: States, societies, and the fate of liberty*. Penguin Press.
Acemoglu, Daron, Autor, David H., Hazell, Jonathon, & Restrepo, Pascual. (2021). AI and jobs: Evidence from online vacancies. NBER Working Paper No. 28257.
Acemoglu, Daron, Kleinberg, Jon, & Mullainathan, Sendhil. (2022). Problems of online communication. Work in progress.
Acemoglu, Daron, Makhdoumi, Ali, Malekian, Azarakhsh, & Ozdaglar, Asu. (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Micro 14*(4), 218–256.
Acemoglu, Daron, Makhdoumi, Ali, Malekian, Azarakhsh, & Ozdaglar, Asu. (2023). A model of behavioral manipulation. Work in progress.
Acemoglu, Daron, Manera, Andrea, & Restrepo, Pascual. (2020). Does the US tax code favor automation? *Brookings Papers on Economic Activity 2020*(1), 231–285.
Acemoglu, Daron, Ozdaglar, Asu, & Siderius, James. (2021). Misinformation: Strategic sharing, homophily and endogenous echo chambers. NBER Working Paper No. 28884.
Agarwal, Ajay, Gans, Joshua S., & Goldfarb, Avi. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.
Allcott, Hunt, & Gentzkow, Matthew. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 211–236.
Applebaum, Ann. (2020). *Twilight of democracy: The seductive lure of authoritarianism*. Signal.
Athey, Susan, Catalini, C., & Tucker, Catherine. (2017). The digital privacy paradox: Small money, small costs, small talk. NBER Working Paper No. 23488.
Autor, David H. (2014). Skills, education and the rise of earnings inequality among the other 99 percent. *Science 344*(6186), 843–851.
Autor, David H., Dorn, David, & Hanson, Gordon H. (2013). The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review 103*(6), 2121–2168.

Autor, David H., Levy, Frank, & Murnane, Richard J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics 118*(4): 1279–1333.

Bergemann, Dirk, Bonatti, Alessandro, & Gan, Tan. (2021). *Markets for information*. Yale mimeo.

Boix, Carles. (2024). AI and the economic and informational foundations of democracy. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Bostrom, Nick. (2014). *Superintelligence*. Danod.

Brynjolfsson, Erik, & McAfee, Andrew. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Budd, Christopher, Harris, Christopher, & Vickers, John. (1993). A model of the evolution of duopoly: Does the asymmetry between firms tend to increase or decrease? *Review of Economic Studies 60*(3), 543–573.

Choi, J. P., Jeon, D.-S., & Kim, B.-C. (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics 173*, 113–124.

Christian, Brian. (2019). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Farboodi, Maryam, Mihet, R., Philippon, Thomas, & Veldkamp, Laura. (2019). Big data and firm dynamics. *American Economic Review: Papers and Proceedings 109*, 38–42.

Ford, Martin. (2015). *Rise of robots*. Basic Books.

Guriev, Sergei, Henry, Emeric, & Zhuravskaya, Ekaterina. (2022). Checking and sharing alt-facts. *American Economic Journal: Policy 14(3), 55–86*.

Hanson, Jon D., & Kysar, Douglas A. (1999). Taking behavioralism seriously: Some evidence of market manipulation. *New York University Law Review*, *74*, 630.

Huntington, Samuel P. (1991). *The third wave: Democratization in the late twentieth century*. University of Oklahoma Press.

Jones, Charles I., & Tonetti, Christopher. (2020). Non-rivalry in the economics of data. *American Economic Review 110*(9), 2819–2858.

Judis, John B. (2016). *The populist explosion: How the great recession transformed American and European politics*. Columbia Global Reports.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics 133*(1), 237–293.

Klemperer, Paul. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics and international trade. *Review of Economic Studies 62*(4), 515–539.

Klinova, Katya. (2024). Governing AI to advance shared prosperity. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Korinek, Anton, & Balwit, Avital. (2024). Aligned with whom? Direct and social goals for AI systems. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Lanier, Jaron. (2018). *Ten arguments for deleting your social media accounts right now*. Hoffmann and Co.

Lee, David S. (1999). Wage inequality in the United States during the 1980s: Rising dispersion or falling minimum wage? *Quarterly Journal of Economics 114*(3), 977–1023.

Lee, Kai-Fu. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Levitsky, Steven, & Ziblatt, Daniel. (2018). *How democracies die*. Crown.
Levy, Ro'ee. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review 111*(3), 831–870.
MacCarthy, M. (2011). New directions in privacy: Disclosure, unfairness and externalities. *I/S 425*(*2011*). https://ssrn.com/abstract=3093301.
Marantz, Andrew. (2020). *Antisocial: Online extremists, techno-utopians and the hi-jacking of the American conversation*. Penguin.
Markoff, John. (1996). *Waves of democracy: Social movements and political change*. Pine Forge Press.
Mishra, P. (2017). *Age of anger: A history of the present*. Macmillan.
Mosleh, Mohsen, Martel, Cameron, Eckles, Dean, & Rand, David G. (2021). Shared partisanship dramatically increases the social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences 118*(7), 1–3.
Neapolitan, Richard E., & Jiang, Xia. (2018). *Artificial intelligence: With an introduction to machine learning*. 2nd ed. Chapman and Hall/CRC.
Pasquale, Frank. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
Posner, Eric A. (2002). *Law and social norms*. Harvard University Press.
Posner, Eric A., & Weyl, E. Glen. (2019). *Radical markets*. Princeton University Press.
Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Press.
Russell, Stuart J., & Norvig, Peter. (2009). *Artificial intelligence: A modern approach*. 3rd ed. Prentice Hall.
Sandel, Michael J. (2020). *The tyranny of merit: What's become of the common good?* Penguin Press.
Shapiro, Carl, & Stiglitz, Joseph E. (1984). Equilibrium unemployment as worker discipline device. *American Economic Review 74*(3), 433–444.
Simonite, Tom. (2020). Algorithms were supposed to fix the bail system. They haven't. *Wired*. https://www.wired.com/story/algorithms-supposed-fix-bail-system-they-havent.
Snyder, Timothy. (2017). *On tyranny: Twenty lessons from the twentieth century*. Tim Duggan Books.
Sunstein, Cass. (2001). *Republic.com*. Princeton University Press.
Thompson, Derek. (2019). Should we be afraid of AI in the criminal-justice system? *The Atlantic*. https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/.
Tirole, Jean. (1989). *Industrial organization*. MIT Press.
Tirole, Jean. (2021). Digital dystopia. *American Economic Review 111*(6), 2007–2048.
Varian, Hal R. (2009). Economic aspects of personal privacy. In W. Lehr, and L. Pupillo (Eds.), *Internet Policy and Economics* (pp. 101–109). Springer. https://doi.org/10.1007/b104899_7.
Vosoughi, Soroush, Roy, Deb, & Aral, Sinan. (2018). The spread of true and false news online. *Science 359*, 1146–1151.
West, Darrell M. (2018). *The future of work: Robots, AI and automation*. Brookings Institution Press.
Zuboff, Shoshana. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

CHAPTER 35

AI AND THE ECONOMIC AND INFORMATIONAL FOUNDATIONS OF DEMOCRACY

CARLES BOIX

Technological innovation has historically shaped the way we have organized and managed our collective and political life. Industrialization, both by changing the patterns of human habitat and the structure of employment and by raising living standards, facilitated the birth of mass politics—and, arguably, modern representative democracy. Successive military technologies, from metal arms to chemical weapons, have shifted the balance of power between the state and its citizens. New means of communication, such as the printing press, newspapers, and the radio, spawned the formation of a critical public opinion capable of holding policy-makers accountable. Artificial intelligence (AI) should be no different. By both fostering the automation of production activities and speeding up the collection, processing, and distribution of information, AI's development should affect our governing institutions as well.

In this chapter, I assess AI's potential impact on the future of democracy as follows. After offering a succinct discussion of the foundations of democracy in Section 1, I explore the mechanisms or channels through which AI may reinforce or disrupt these foundations. Sections 2 and 3 describe two fundamental economic transformations brought about by AI that may affect politics: a shift in the nature of employment resulting in higher income inequality and the potential concentration of capital ownership with a corresponding decline in market competition. Section 4 considers, in turn, what those changes may imply for democracy in both developed and developing countries. In advanced democracies, AI may lead to political polarization due to a growing divide between capital and high-skilled individuals on the one hand and the rest of labor on the other. Nonetheless, that development may not necessarily erode democratic institutions thanks to the "re-equilibrating" role of elections—the process through which electoral competition incentivizes policy-makers to respond to new social and economic challenges. By contrast, in emerging and peripheral economies, particularly those with weak or no democratic mechanisms, AI's

effects are more likely to be de-democratizing. Section 5 explores the contradictory effects of AI on the generation, collection, and use of information, and, as a result, on politics. On one hand, digital technologies have been celebrated for their capacity to break information monopolies of media conglomerates, strengthen political accountability, and facilitate the coordination of opposition movements against authoritarian regimes. On the other hand, they have been blamed for fragmenting and polarizing public opinion in democracies and for supplying authoritarian regimes (and even democratic states and big firms) with powerful surveillance tools. Section 6 concludes by discussing a set of interventions to strengthen democratic institutions.[1]

Foundations of Democracy

To understand the potential effects of AI on democracy, let us start by defining the latter as a procedure in which its citizens decide, by casting a vote or a sequence of votes, how to govern themselves—that is, what rules should bind their collective life, what should be the optimal distribution of assets, and so on. This procedure implies two things. First, a majority of the population determines the position (or welfare) of each member of the population and, therefore, of the minority that has not voted with the majority. Second, voters are, *ex ante*, that is before elections, uncertain about who will win. If they were not, it would mean that the election had not been conducted according to fair and equitable norms; that is, that a section of its participants had manipulated the voting procedure to their advantage.

Given those two key conditions, democracy becomes possible only when voters accept (the possibility of) losing elections and, as a result, policy outcomes different from their preferred alternative. Now, acquiescing to a democratic system will depend on the value voters derive from it as opposed to an authoritarian regime. If the expected gain from submitting themselves to elections is larger than the net benefit they would accrue from supporting a nondemocratic regime, they will accept democracy. Otherwise, they will not.

Those calculations depend, in turn, on two main variables: the structure of social interests and the costs of imposing a non-democratic regime. On the one hand, the likelihood of an authoritarian solution will rise with the heterogeneity of interests in a given political community. The losing party or section in any (potential or already realized) election will be less supportive of democracy the higher the difference between its preferred policies and the ones approved by a democratic majority.[2]

Although that heterogeneity may have different roots, the structure of the economy plays a crucial role in shaping it in three ways. First, rising income (and wealth) inequality tends to exacerbate redistributive tensions, moving the majority to raise taxes and transfers to curb that inequality. High-income individuals, who generally constitute a minority in society, may then become less inclined to accept a democratic regime. Second, economic growth tempers the redistributive impact of inequality and therefore the (high-earners) resistance to democratic rule. If the marginal utility of income declines when income rises, then, as economic development takes place, wealthier individuals will be more willing to tolerate higher taxes and therefore to put up with the consequences of democracy.[3] Finally, asset holders will be more inclined to accept democracy when the value of their wealth does not depend on electoral results; that is, when governments cannot regulate its price and/

or return. That is particularly the case for assets that can escape from state monitoring or be moved to other jurisdictions in response to any regulatory actions—or, to employ more technical terms, when the assets are not specific to the place where they are exploited.[4] In Sections 2 and 3, I explore how AI changes these parameters (income distribution, level of income, type of assets) and therefore the likelihood of democracy.

On the other hand, imposing an authoritarian solution depends on the costs incurred to exclude part of the population from voting—either through fraud or, directly, by disenfranchising them. Those costs have historically varied with the technologies available to both government (to enforce that exclusion) and opposition (to challenge it). As I discuss in more detail in Section 4, neither the recent revolution in information and communication technologies nor the potential applications of AI are exceptions to that rule.

AI and Labor

As with any technological innovation, AI has the potential to change the economy's production function; that is, the relative contributions of capital and labor to the generation of output, the corresponding returns to each of those factors, and, therefore, the political incentives of their holders toward democracy. In this section, I consider the impact of AI on labor. In the following section, I turn to its potential effects on capital.

Technological change and shifting capital–labor complementarities

Production technologies determine the type of labor that is central or, in other words, complementary to capital and the production system in general. In the last 200 years, the demand for labor has switched from demand for the unskilled workers that toiled in the first Manchester factories at the time of the first industrial revolution to the highly educated employees of today's Silicon Valley companies—with consequential effects on the overall structure of employment and distribution of income.

During the first industrial revolution, unskilled factory workers replaced artisans working in small workshops as the main kind of labor employed in a growing manufacturing sector. Their low wages, the declining standards of living in the new industrial towns, at least until the last decades of the nineteenth century, and the rising profits of capital heightened inequality and social conflict, making full democracy impossible. Voting rights were circumscribed, at most, to well-to-do men.

The introduction of the assembly line and mass production techniques, as well as the use of electricity and electric motors at the turn of the twentieth century, transformed again the structure of production and employment. Semi-skilled and skilled employees replaced unskilled workers as the main type of labor complementary to capital. Riding on rapid economic growth, wages rose across the board, particularly among middle social strata, making the distribution of earnings more equal. The Gini coefficient, which was around 0.5 or higher in the late nineteenth century, declined throughout the middle

decades of the twentieth century to about 0.3 in North Atlantic economies. In the wake of higher salaries and living standards and a more equal income distribution, social conflict declined and democratic rule became fully entrenched in all advanced economies by the middle of the twentieth century. In the postwar period, 75 percent of all countries in the world's top quintile in levels of development were democracies. Only nine percent of countries in the bottom two quintiles were democratic. Democracy came with consensual politics and moderate electoral platforms. Conservative and socialist parties, for a long time at odds with each other, embraced the key tenets of what many have labeled "embedded liberalism" or "democratic capitalism": free and fair elections, competitive markets, and a welfare state that mitigated the risks of economic downturns and protected citizens from illness and aging.

The relationship between democracy and twentieth-century capitalism became, in fact, a two-way, symbiotic one. Economic growth (and the particular production structure of the second industrial revolution) led to social peace and liberal institutions. Yet, at the same time, democracy, and the kind of welfare state it spawned, reinforced growth and its equitable distribution. Business benefited from having a well-trained, healthy labor force. In turn, voters saw markets as an efficient system to generate growth and to give them better life chances (on the basis of effort and merit). Indeed, to support that economic system, public opinion favored a regulatory framework aimed at reinforcing, even if imperfectly, competitive markets. Antitrust laws led to the break-up of monopolistic structures. A restrictive patent system attenuated the consolidation of first movers in a particular economic sector, fostering a continuous process of technological innovation.

The invention of the personal computer and, later on, the internet, email, and mobile phones transformed again the structure of the production process and, with it, of employment. Automation accelerated in the manufacturing sector, leading to a collapse in its employment numbers. In Europe, for example, manufacturing jobs represented over one-fifth of all employment in 1970 but less than one-tenth in the middle of the 2010s. More importantly, automation spread to nonmanual routine jobs through the use of software programs that reproduced a growing set of administrative tasks in a wide range of traditional white-collar jobs, from accounting and banking to travel agencies. Whereas almost 45 percent of the working-age population in the United States worked in routine occupations in the mid-1980s, only 31 percent did in 2014 (Cortes et al., 2016). In the meantime, professional and managerial jobs highly reliant in abstract, relatively creative thought processes, rose steadily. In the United States, the share of high-skill occupations (managers and professionals) over total employment grew from almost 28 percent of all civilian employment in 1980 to 39 percent in 2010 (Katz & Margo, 2014). In short, highly educated workers became the main type of labor complementary to capital.

A changing labor market came hand in hand with a shifting wage and income distribution. Labor productivity and median earnings, which had trended together until 1975, diverged afterward. U.S. labor productivity doubled between 1975 and 2016. By contrast, median earnings remained flat throughout the whole period. In economies with flexible labor markets (mostly Anglo-American countries), wage and income distribution broadened. Wages for U.S. workers dropped in real terms for individuals in the bottom quintile of the earnings distribution and stagnated for those around the median while doubling for individuals with postgraduate education.[5] By contrast, in highly regulated economies (mostly continental Europe), low wages rose and earnings inequality remained

unchanged—but the cost was minimal job growth or even a fall in jobs in the private sector in net terms.

AI and quasi-automation

Although predictions about future trends in the process of capital–labor substitution due to technological innovation are highly uncertain, we may consider two possible scenarios under AI. In the first one, which we may want to refer to as "quasi-automation," AI substitutes all kinds of labor except for very high-skilled occupations. In the second one, AI replaces labor in all production and distribution tasks, resulting in a system of "full automation."

Quasi-automation, where AI makes unskilled and semiskilled labor superfluous while high-skilled workers are still needed to produce goods and services, seems at this point the most probable outcome—with the caveat that we do not know *when* it may come to pass. Frey and Osborne (2017) conclude that "transportation and logistic occupations, together with the bulk of office and administrative support workers, and labour in production occupations" (p. 265) have a high probability or risk of computerization. By contrast, "generalist occupations requiring knowledge of human heuristics, and specialist occupations involving the development of novel ideas and artifacts" (Frey & Osborne, 2017, p. 266) have a much lower one.[6] Vulnerability to computerization is correlated with the level of skills—extremely high "for low-skill and low-wage jobs in the near future" and much less so for "high-skill and high-wage occupations" (p. 267).

The consequences of that change will be arguably different for advanced economies, new industrializing countries, and developing nations. Coinciding with the information and computational revolution that started in the 1970s, and in part fostered by it, the process of economic globalization implied a critical transformation in the international division of labor. As transportation and communication costs declined, a subset of developing countries were able to use their unskilled labor force to build up a manufacturing base. In addition, multinational corporations unbundled their production structure to exploit the specific comparative advantage of each country across the world—mostly maintaining operations based on highly qualified employees in the northern core while moving or subcontracting tasks performed by less skilled labor to the developing world.[7]

Under a scenario of quasi-automation, however, subcontracting and/or moving low-skill tasks to emerging economies may become pointless. With no comparative advantage to be drawn from placing fully robotized plants in periphery economies whose main attraction is having cheap labor, factories could be located anywhere in the world. That could lead businesses to "re-shore" production back to the consumer markets of advanced economies to minimize distribution and transportation costs.

In advanced economies, AI will increase productivity, total output, and per capita income by saving on labor. Its internal redistributive effects will depend, however, on the capacity of individuals to acquire the relatively high level of abilities demanded by the production model of the future. If education investments are enough to move individuals upwards in the skill ladder (and there are enough jobs available in high-skilled sectors), social conflict will remain subdued. If not, a fraction of the workforce will remain unemployed or employed but lowly paid, contributing to persistent inequality.

In emerging economies, in turn, the impact of quasi-automation will be mediated by the costs of moving up in the production ladder from low-value-added to high-value-added activities. If the barriers to capital investment or innovation are low, those economies will experience continuous economic growth. However, if those barriers are high, catching up with Europe, Japan, and the United States may never happen. Notice, however, that even in the more optimistic case, middle-income economies may experience, first, some economic backsliding because the process of production "re-shoring" will deprive them of current foreign markets, and, second, the same kind of low-skilled/high-skilled tensions taking place in advanced economies.

Finally, low-income countries are likely to remain stagnant. Lacking in any industrial manufacturing basis, which has historically been the stepping-stone toward development, it may be hard for their very unskilled labor force to acquire enough skills and technical capacities to become complementary to the technologies of the future. They may therefore keep being the periphery of the periphery—providing primary products to the rest of the world and services such as tourism and so on.

Full automation

Under the more radical scenario of full automation, no labor will be employed at all at any stage of production. As in the case of quasi-automation, firms will locate their production close to their consumers. Advanced industrial economies will "reindustrialize" (without creating any new jobs) to serve their populations. Here the distributional split will not take place between capitalist and high-skilled individuals on one side and the rest of labor on the other, but just between capital and labor.[8]

In middle-income and poor countries, the outcome will be a function of the cost of capital. If innovation and capital investment remain costly, the process of economic modernization of middle-income countries may stop and even reverse itself. The economic take-off of poor countries may be completely out of the question. If, however, technological innovation reduces the cost of capital (as has happened for cell phones and laptops), even poor countries may have a chance to expand their manufacturing sector (mostly for internal consumption). Still, the internal distribution of gains and losses will follow the same pattern of advanced countries.

AI and Capital

AI may affect capital in two ways: through its contribution (relative to labor) to production (and therefore its share of total income) and through capital's ownership structure.

Relative weight of capital

The relative weight of capital in the production process will again depend on the extent to which AI either complements and enhances (high-skilled) labor or results in full

capital–labor substitution. In the first instance (of capital-high-skills complementarity), the owners of AI (or those with access to funds to invest in them) will robotize production, appropriating a larger share of the value of the unit or service produced, as soon as the price of using the new technology falls relative to the wage paid to (low-skilled) workers. In the case of full automation, only one factor—capital—will be employed and its owners (i.e., capitalists) will appropriate all returns from production.

There are already some signs that capital has grown in relative terms within the economy. Among new firms, the ratio of market value to number of employees has exploded. In 2020, the number of employees stood at 147,000 for Apple, 135,000 for Google, 163,000 for Microsoft, and 58,000 for Facebook. GM, one of the quintessential companies of twentieth-century capitalism, employed, at its maximum, 97,000 individuals in the United States. By early 2021, Apple, Google, Microsoft, and Facebook had a market capitalization ranging from over $700 billion to $1,715 billion—15 to 40 times bigger than GM at the peak of its market value. In the whole American economy, capital income grew at an annual rate of 2.2 percent while labor income barely inched up at a rate of 0.1 percent per year between 2000 and 2014. The U.S. labor share of national income fell from about 64 percent throughout the postwar period to 58 percent from the mid-1980s onwards (Elsby et al., 2013). Similarly, Karabarbounis and Neiman (2014) report a fall of five percentage points in the labor's income share in a sample of 59 countries in the period from 1975 to 2013. By economic sectors in the United States, the labor share of income fell rapidly in sectors with high R&D intensity—from 80 percent in the mid-1970s to 60 percent in 2011. By contrast, it did not vary in less R&D-intense sectors (Guellec & Paunov, 2017).[9]

Asset ownership

The impact of AI on the concentration of capital is harder to predict—and could go either way depending on how it may affect, first, the costs of innovation and, second, information costs.

If the barriers to AI innovation and implementation are high, the ownership of capital will concentrate in the hands of "a limited number of exceedingly wealthy property owners," to use the words of Nobel laureate James E. Meade back in 1964. AI may lower, however, the costs of producing machines (i.e., of capital investment) directly—to the point of enabling everybody to set up some heavily automatized or robotized shop. Mobile phones have already eased African farmers' access to prices, weather conditions, and state-of-the-art agricultural techniques. Digital platforms have made the use of time by independent truck drivers more efficient. Craft shops and local tour operators in developing countries can contract directly with customers in advanced economies.

The information effects of AI can go both ways too. AI technologies may foster the rise of larger firms and of oligopolistic markets by minimizing the costs of both collecting information about (present *and* future) preferences of buyers/consumers and integrating production chains. Currently, for example, Google has over 90 percent of the search engine market, Facebook controls almost 70 percent of social networks, and three top firms concentrate almost one-third of all the U.S. e-commerce. On the other hand, AI may

push down the costs of monitoring production across tasks, weakening the incentives to integrate all jobs in a single plant or under a single firm. Instead of the large corporations of twentieth-century capitalism, there could emerge an economy formed by self-employed individuals engaged in very specific or narrowly defined tasks who then transact with each other in the marketplace. Topcoder has been showcased as a paradigmatic case of this atomized production system: an outsourcing company that offered a worldwide community of around one-and-a-half million freelance programmers, engineers, and developers by early 2021.[10] Likewise, the digital platform Airbnb lets homeowners or small companies offer accommodation online to customers—bypassing the fixed costs incurred by hotel chains and traditional travel agencies. In the limit, that would multiply the number of capital owners (particularly if they have skills complementary to that capital) and lead to a highly fragmented production system similar to the networks of craftsmen that existed in European towns before the industrial revolution.

Political Consequences

Table 35.1 summarizes the economic effects of AI on both capital and labor. Its political consequences will then vary with each country's levels of economic (industrial) development and democratic institutionalization. In emerging and peripheral economies, AI is likely to have a de-democratizing effect. In advanced democracies, its impact will be probably more ambiguous—partly because of the crosscutting consequences of AI and in part because the latter's final effects will be conditional on the policy responses in those countries.

Table 35.1 AI Consequences by Production Factor and the Aggregate Economy

	Quasi-Automation	Full Automation
Factors		
Capital	Gain	Strong gain
	Potential increase in capital concentration	Potential increase in capital concentration
Labor		
Highly Educated	Gain	Loss
Rest of Labor	Loss	Loss
Whole Economy		
Average Income	Higher in advanced economies	Higher in advanced economies
	Unchanged or lower in LDCs	Unchanged or lower in LDCs
Income Dispersion	High	Extremely high & very skewed (if cost of capital is high)
Assets	Assets	Steep decline in specificity

Developed countries

As discussed in Section 2, transitions to and away from democracy vary with three main variables: positively, with level of income; negatively, with income distribution; and, positively, with non-specific assets (i.e., assets that can be deployed in multiple jurisdictions with no significant loss in their value). AI will lead, via productivity gains, to income growth in developed countries, which both have the resources to spur these new technologies and will benefit from the process of production reshoring. AI may also have a pro-democratic effect via the nature of new assets: digital know-how seems to be deployable everywhere and therefore difficult to control across borders. These two "democratic" forces will be checked, however, by a growing gap between capital owners and labor or, at least, between capital owners and high-skilled individuals, on the one hand, and the rest of labor, on the other.[11] The question is, therefore, which of these conflicting forces will prevail.

The likelihood of a full democratic crisis seems low due to both the (democratizing) role of income and the plausible "re-equilibrating" dynamics of democracy. Still, the use of AI may have worrying effects on the quality of democracy.

Income and democracy

A large body of empirical research shows that the probability of having democratic institutions is strongly correlated with per capita income. Whereas over 90 percent of the countries with a per capita income above $10,000 were holding free and fair elections at the beginning of the 21st century, less than one in five countries with a GDP per capita below $2,000 (in constant dollars of 1996) are democratic. Once established, democracy has never died in wealthy countries. Between 1800 and 2007, there were 69 instances of democratic breakdowns; that is, transitions from democracy to dictatorship, such as Germany in 1933 or Chile in 1973. One-half of the democratic breakdowns occurred in countries with a per capita income below $2,500 and one-third in the range between $2,500 and $5,000. No country has reverted to authoritarianism above $10,000.[12]

Nevertheless, if per capita income simply serves as a proxy (in those studies) for the lower income inequality of the twentieth century, we may conclude that an acceleration of today's growing inequality (due to an intensification of technological change) may lead, regardless of whether a country enjoys a relatively high per capita income or not, to the collapse of democracy. Indeed, recent work suggests that current economic and informational changes have resulted in the erosion of democratic institutions across the world (Foa & Mounk, 2016; Levitsky & Ziblatt 2018). Still, this interpretation seems debatable for two reasons. In the first place, poverty has been mostly eradicated in advanced economies. In 1850, about half of the western European population had a per capita income similar (in real terms) to today's poorest countries in Africa. Today, over 90 percent of the population in Europe and North America enjoys an income equal to or higher than the income of an individual in the 95th percentile of the income distribution in those same continents during the first half of the 19th century.[13] This development has, in turn, eliminated what used to be one of the sources, if not the main source, of riots, revolutions, civil wars, and authoritarian coups.[14] In the second place, recent work that relies on the political trajectories of all the sovereign

countries in place since 1900 finds that high levels of development cancel the negative impact of inequality on democratic stability (Beramendi et al., 2021).[15]

Democratic "re-equilibration"

Because democracy can be thought of or understood as a political equilibrium, that is a self-sustaining outcome from which no political player has any incentive to deviate (that is, to challenge it) due to the (economic) calculations described in Section 1, AI can disrupt those incentives to the point of making some actors unwilling to support a democratic regime. Still, and provided the disrupting shock is not too abrupt or large, democracy has a built-in mechanism to manage the shock and sustain the democratic commitments of voters: the electoral process itself.

Spurred by the threat of losing elections, policy-makers respond, in principle, to new political and economic challenges (including those that may jeopardize democratic performance) that harm the electorate (or, at least, the decisive voter). If those new policy interventions are successful, they may reduce economic and social strain, and, therefore, contribute to preserving democracy as a stable political outcome. Policy responses can range from deploying new programs to help workers adapt to technological change (through, for example, new educational investments) to passing new regulations blocking technological innovation. By contrast, this process of re-equilibration, through which democratic procedures stimulate, via elections, policy that brings social conditions back to a pro-democracy equilibrium, is highly unlikely in authoritarian regimes—particularly if its ruling classes are unaffected by the consequences of AI.

Democracy, however, may not be sufficient to deal with the consequences of AI. As stressed earlier, democracy and capitalism reinforced each other throughout the twentieth century: public spending on health and education benefited both voters (because it enabled them to take advantage of social mobility opportunities offered by a booming economy) and business (because it generated the kind of productive labor force needed for mainstream production processes). Depending on its reach, AI could strain that relationship in the future. If AI firms only need to hire a fraction of the labor force, they may resist supporting broad spending and investment programs that are of little use to them. In turn, those voters that have no access to well-paid AI jobs may end up demanding straightforward redistributive programs rather than policies that nurture and complement the new production technologies.

Moreover, democratic re-equilibration can only take place if democracy functions well, that is, if politicians remain accountable to the public, uncaptured by specific individuals or economic sectors. Here lies an additional potential threat of AI. Its owners may employ their rising income and influence to muffle the voice of the majority through campaign contributions, political lobbying, and personal connections. They could also push for legislation to block the entry of any potential rivals and lock in its initial economic advantage.

The growing concentration of wealth that has taken place in the last few decades does not bode well for the future. During the last decades, campaign contributions in U.S. federal elections have gradually become concentrated among the super-wealthy. The top 0.01 percent of households (in the income distribution) donated between 10 and 15 percent of all campaign contributions until the early 1990s. In 2012, the proportion was 40 percent.

The wealthy have displaced other groups such as organized labor as the main source of revenue for political parties. In the 1980s and early 1990s, donations from the top 0.01 percent and from trade unions were roughly similar. In 2012, contributions from the top 0.01 percent were four times bigger than labor's. At least in the United States, money appears to be shaping policy-makers' preferences and votes. According to Larry Bartels, the views of members of Congress are closer to their wealthy constituents than to low-income voters (Bartels, 2008).

The transformation of labor relations in the last decades may reinforce the power of money. Right after the Second World War, at the peak of twentieth-century capitalism, union membership ranged from one-third of all American workers to over two-thirds of the labor force in small European countries. Today, unionization rates stand at around 10 percent in France and the United States, less than one-fifth in Germany and around one-fourth in the United Kingdom.

Developing countries

For less developed countries, the political horizon looks darker. As in advanced economies, they may experience increased inequality (due to AI). However, without the financial resources and state capacity of richer countries, those effects may result in stronger social and political tensions. In fact, if automation and reshoring intensify, industrializing countries may be unable to escape from a middle-income trap. In the limit, they may even suffer from some form of economic backsliding—negative growth rates and a shrinking middle class. As a result, democratic institutions and practices will become more fragile in existing democracies. In authoritarian polities, the likelihood of a democratic transition may decline.

AI and the Costs of Exclusion

The mechanisms employed to suppress (and, conversely, to support) democracy are of two kinds. On the one hand, the likelihood of democracy depends on the cost of the specific tools that states deploy to suppress dissent through force: the higher the power differential between governments and citizens, the easier it will be for an incumbent to exclude the opposition.

On the other hand, the likelihood of democracy depends on the type of information that both opposition groups and governments have about each other. Information plays a key role among opposition groups in the following sense. Demonstrations, protests, and social mobilization in general rely on the coordination of a sufficiently high number of individuals to be successful. In turn, that coordination is only possible if potential demonstrators (or voters) know (or at least expect with some certainty) that other individuals like them will mobilize too against the incumbent or the regime (Kuran, 1991).

Information about their citizens is also essential to governments, and, particularly, authoritarian governments, in two ways. The first function of information consists in facilitating the management of repression. A central problem of dictators is that they

operate, by definition, in an information-poor environment. A dictator does not know whether the support he receives is sincere or purely strategic (and, in the latter instance, likely to collapse once his subordinates calculate they can get rid of him successfully). Better knowledge about the preferences of their agents and, more generally, of citizens, and about the political actions they engage (or plan to engage) in allows dictators to act preemptively against any coups, revolts, and other actions of resistance. It also enables them to repress their opponents in a more targeted fashion. Targeted repression is less costly than indiscriminate repression, less likely to result in international sanctions, less prone to generate a backlash among the population, and, in general, more effective (Siegel, 2011; Guriev & Treisman, 2020; Xu, 2020). Information plays a second function too: generating political consent. The control and definition of news and opinion through either standard or social media helps regimes to distort the beliefs of the opposition and to shape public opinion in general.

AI may affect the type of repressive tools of incumbents (and of resistance tools of oppositions). For example, at some point, the former may count on a completely "robotized" police force. However, AI's effects are likely to be particularly strong in the domain of information collection and use.

On the one hand, AI may strengthen opposition movements—mainly by easing their costs of coordination and by generating news and opinion that weaken the legitimacy of and support for authoritarian regimes. Diamond (2010) famously hailed the internet as a "liberation technology" that would enable "citizens to report news, expose wrongdoing, express opinions, mobilize protest, monitor elections, scrutinize government, deepen participation, and expand the horizons of freedom" (p. 70).[16] Still, the existing empirical research has found little evidence that digital technologies have had any impact on the capacity of anti-regime movements to democratize authoritarian systems (Lynch, 2011; Gunitsky, 2015).

On the other hand, AI may empower (authoritarian) states through three main channels. In the first place, political incumbents may (and, in some cases, already do) employ these new technologies to engage in the "soft" identification of their opponents. The Chinese state monitors the interactions that occur in the country's social media to extract information about the interests and grievances of its citizens and to tailor policy accordingly (Qin et al., 2017; Huang et al., 2019). Referring to Putin's rule, opposition leader Alexei Navalny declared back in 2010 that "Internet for the government is some kind of a focus group. [The government] just like[s] to do what the people want. I mean, if it doesn't contradict their own interests. The political agenda, however, will be tested on the Internet."[17]

In the second place, AI may facilitate the "hard" identification and control of opponents. As Carly Nyst at Privacy International writes, "in the digital era, almost every secret we keep, we keep online, whether we are aware of it or not" (Nyst, 2018. p. 12). The construction of a digital profile for each individual makes surveillance, which in the past relied on physical monitoring and was constrained by storage and information-processing capabilities, much easier to finance and implement. Digital surveillance can then be combined with direct physical surveillance, as is the case with China. There, public authorities approved China's Golden Shield Project in 1998 to build population databases, ID tracking systems, and internet surveillance tools. A few years later, they started to integrate them with street surveillance cameras and facial recognition techniques. By 2017, China's network included 176 million surveillance cameras—with many of them having the capability of transforming some traits of passersby into data. That same year, China launched pilot programs to detect

particular voices on phone conversations automatically (Qian, 2019). As AI speeds up the collection and processing of data, the capacity of governments to detect and preempt their opponents may become formidable.

Finally, AI's effects may go beyond the identification, cooptation, and/or control of citizens. States are employing (and will employ) digital technologies to shape public discourse directly: discrediting the opposition through fake news, disseminating favorable news about the achievements of the regime in a systematic way (often delivered by supposedly independent agents), censoring partially and strategically the blogosphere, and mobilizing the regime's supporters in favor of the government and against its adversaries (Gunitsky, 2015).

The mostly pro-incumbent bias of AI means that its political impact is likely to vary with the institutional point of departure of each country. In those polities endowed with robust democratic institutions (i.e., having a working constitutional structure and a strong civil society), we may expect legislation and institutional structures to preempt the abusive use of AI from governments and large private actors. By contrast, in authoritarian regimes, AI may simply reinforce the position of incumbents.[18]

Interventions

Several policy interventions seem advisable to minimize the negative effects of AI on democracy:

1. Speeding up the transition of the workforce from routine jobs, which are at the highest risk of automation, to nonroutine occupations or, at least, jobs that are complementary to AI tasks. The key intervention would consist in investing in an educational system both extensively (in terms of the number of people targeted) and intensively (employing educational strategies such as, perhaps, active learning, that prepare individuals for flexible, creative jobs).
2. Pre-empting the formation of a closed elite. That requires maintaining competitive markets—mostly through an active antitrust policy. It may also imply legal reforms to strengthen democracy of the kind listed in the following point.
3. Implementing reforms to strengthen the representation of the common voter by limiting campaign donations by corporations, democratizing the distribution of electoral funds along the lines of the reform proposed by Bruce Ackerman and Ian Ayres,[19] and disclosing the (ownership and marketing) relations between media and large firms.
4. A growing number of voices have called for the extension of a universal basic income. However, its implementation is not cost-free. First, according to pre-trials in Canada and the United States, it reduces the incentives people have to work (Sage & Diamond, 2017). Second, it may subsidize the wages paid by firms, freeing them to offer lower wages, and therefore redistributing resources to employers. Third, it may reduce individuals' incentives to school themselves. Fourth, it may lead to a backlash from voters that have permanent jobs—successful welfare states rely on the idea of risk sharing rather than on pure redistributive schemes.

A different strategy would consist of establishing a "universal basic capital" (UBC); that is, granting each person some fixed capital at birth (in line with a proposal already made by Thomas Paine at the end of the 18th century). To reward prudential behavior and avoid the possibility of their owners to squander it, its recipients would have free disposition only over its returns (and only after becoming legally adult). That solution would combine the supporting component of a universal basic income with an individual incentive to manage it actively; that is, to put some effort instead of consuming an income flow passively. The UBC could be funded through a tax on robots. In any case, its implementation seems advisable if and after AI reduces the number of available jobs and creates a permanently unemployed class.

5. If AI ends up harming less developed economies (through the process of automation and reshoring) and intensifies interregional inequalities (reversing the trend of the last decades toward cross-country economic convergence), establishing a system of open borders that allows labor migration to the North. That system may become the main way for people in the South to escape poverty; that is, it may be the only redistributive tool in our hands to equalize life chances across the world. Given the political backlash such an open-border system may create, especially if AI intensifies a skill divide, migration should probably be gradual.[20]
6. Establishing independent boards or agencies to supervise the use of information by states (and corporations) and passing stringent laws to preserve the privacy of citizens (and to minimize the use and manipulation of "digital profiles").
7. Developing an international framework to maximize the free flow of information and to break through the control enlercised by authoritarian states. That strategy would include the use of satellite-based information transmission systems, stronger incentives to establish digital communications through peer-to-peer networks, and so on.

Conclusions

The effects of AI on democracy are not just complex but also partly conditional on the current operation of our democratic institutions. To survive, democracies require that voters accept them and the possibility of losing elections. Relying on a long line of research, this chapter has outlined the conditions that facilitate a generalized acquiescence to democracy: relative high levels of development, moderate to high equality of conditions, the protection of valuable assets against the possibility of expropriation, and a balanced distribution of capabilities and informational resources between incumbents and oppositions.

AI may disrupt some or many of those conditions. The process of capital–labor substitution it implies may generate more inequality and, in fact, impoverish parts of labor. Those effects may happen both in already developed economies—in fact, they are already experiencing growing economic polarization—and in developing countries. It is indeed likely that, in the long run, AI may be more disruptive in the latter. From an informational point of view, it is also likely that AI may strengthen the resources of governments and incumbents, particularly if they already rely on authoritarian institutions.

At the end of the day, however, AI should not be seen as operating as a "deus ex machina." Its effects will depend on the political responses we devise. Facilitating the education of labor should allow most of the population to benefit from AI. Establishing the proper kind of institutional and regulatory guarantees should make sure that governments do not exploit the increasing information and control capacities that AI appears to be generating.

Notes

1. Due to space considerations and thematic coherence, this chapter does not cover two additional topics that connect democracy and AI. First, the impact of social media on public opinion, in terms of misinformation, preference polarization, social segmentation, etc. Jungherr et al. (2020) offer a recent and thorough review on this question. Second, a growing literature explores the impact of automation on electoral preferences as a result of employment losses and wage adjustment; see Anelli et al. (2019), Thewissen & Rueda (2019), and Kurer (2020).
2. On the foundations of democracy from a theoretical point of view, see Dahl (2008), Przeworski (1991, chapter 1), Weingast (1997), Boix (2003), and a recent review in Svolik (2019). Recent empirical work on this questions includes Przeworski et al. (2000), Boix (2011), Miller (2012), and Treisman (2015).
3. Recent work on the relationship between income and life satisfaction has found that even though higher incomes are associated with more happiness—with the former driving the latter in those cases where individuals get richer by chance (i.e., through lotteries), greater wealth exhibits a diminishing positive correlation with life satisfaction (Layard et al., 2008; Frey, 2010).
4. For an extensive formal discussion of these conditions, see Boix (2003) and, more recently, Beramendi et al. (2021).
5. On a changing labor market, see Autor (2010, 2015) for the United States, and Goos et al. (2014) for Europe.
6. According to the World Development Report of 2016, two-thirds of jobs in developing countries and between 50 and 60 percent in Europe and the United States could be automated over the coming decades (World Bank, 2016, p. 126). Employing different criteria may lead, however, to sharply different results. Arntz et al. (2016) estimate that only nine percent of jobs in OECD countries are highly automatable.
7. On changes in transportation costs, see Bernhofen et al. (2016) and Hummels & Schaur (2013). On outsourcing, see, among a broad literature, Baldwin (2016) in general, and, for specific regions, Hanson et al. (2005), Bayard et al. (2015), Ando & Kimara (2011), and Kinkel et al. (2007).
8. The nature of these distributional effects will likely vary by sector. Some sectors will experience a capital/labor split. Others may continue to be defined by a cleavage between capital and skilled labor on one side and the rest on the other.
9. On the other hand, returns to capital may decline over time if information and communication technologies raise inter-firm competition. The introduction of digital technologies has reduced the costs of advertisement and distribution. For purely digital products, transportation costs are now zero, therefore reducing part of the high barriers to entry that producers of physical goods faced in the past. In a global survey conducted in 2015, more than two-thirds of firms in the non-digital economy stated that they were experiencing higher levels of competition due to digital technologies (World Bank, 2016).

10. See Malone et al. (2011). Information from https://www.topcoder.com/, as of March 13, 2021.
11. AI may replace the decentralized information-generation system provided by markets, theoretically allowing the emergence of an efficient planner. If so, the state could replace the market, leading to a vast rearrangement of income flows and asset ownership. Control of the state would become crucial, strongly raising the stakes of elections.
12. The data come from Boix et al. (2013). For a discussion of the (extremely large) literature on democratization and democratic stability, see Geddes (2007) and Boix (2011).
13. Own calculation based on the data from Bourguignon & Morrisson (2002).
14. See, for example, Londregan & Poole (1990) and Miguel et al. (2004).
15. In addition, higher inequality may not result in social strife in affluent societies provided this inequality is rooted in processes that are (or that public opinion believe to be) fair and equitable. Although public concern for current (and rising) levels of inequality has risen, survey respondents seem to prefer some inequality rather than complete equality when they are asked about the optimal distribution of income (Norton & Ariely, 2011; Kiatpongsan & Norton, 2014). Public opinion tolerance for inequality seems to be conditional on the effective operation of some principle of fairness and moral desert (Almås et al., 2020; Starmans et al., 2017).
16. See also Tufekci (2017) and, for a detailed study of digital technologies in the context of urban revolutions, Beissinger (2021).
17. Quoted in Gunitsky (2015, p. 48)
18. Although not pursued here, AI may affect the kinds of strategies of resistance and protest among opposition groups. Physical protest, unless it involves a spontaneous, sudden, massive movement against the regime, will be less likely, precisely because governments will have an increasing number of tools to control and preempt the coordination of their adversaries. By contrast, political opposition may take the form of disruptions actions against AI systems ("cyberattacks," broadly construed).
19. See Ackerman & Ayres (2008). The Ackerman and Ayres' system consists in giving to each citizen a fixed number of dollars to be spent in the electoral campaign in the way (that is, on the candidate) everyone prefers. That proposal is complemented with the decision to establish a blind trust in which all private donations are put—to be transferred to the candidate or parties chosen by the donor. As with the secret ballot, the secrecy of donations may reduce the lobbying by well-identified donors.
20. Regardless of the impact of automation, due to the current demographic trends in advanced economies (with current below-replacement fertility and negative natural population growth in the near future), a migratory flow from South to North seems fully sustainable and in fact economically advisable.

References

Ackerman, Bruce, & Ayres, Ian. (2008). *Voting with dollars: A new paradigm for campaign finance*. Yale University Press.

Almås, I., Cappelen, A. W., and Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians? *Journal of Political Economy* 128(5), 1753–1788.

Ando, Mitsuyo, and Fukunari Kimura. 2011. *Globalizing Corporate Activities in East Asia and Impact on Domestic Operations: Further Evidence from Japanese Manufacturing Firms*. Research Institute of Economy, Trade and Industry (RIETI), Discussion Paper 11-E-034. Tokyo: RIETI.

Anelli, M., Colantone, I., & Stanig, P. (2019). We were the robots: Automation and voting behavior in western Europe. *BAFFI CAREFIN Centre Research Paper*, (2019–115).

Arntz, M., Gregory, T., and Zierahn, U. (2016). The risk of automation for jobs in OECD countries: A comparative analysis. *OECD Social, Employment and Migration Working Papers*, No. 189. OECD.

Autor, D. H. (2010). The polarization of job opportunities in the US labor market: Implications for employment and earnings. *Center for American Progress and The Hamilton Project 6* , 11–19.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives 29* (3), 3–30.

Baldwin, R. (2016). *The great convergence: Information technology and the new globalization*. Belknap Press.

Bartels, Larry M. (2008). *Unequal democracy: The political economy of the new gilded age*. Princeton University Press.

Bayard, K., Byrne, D., and Smith, D. (2015). The scope of US factoryless manufacturing. In S. N. Houseman and M. Mandel (Eds.), *Measuring globalization: Better trade statistics for better policy*, vol. 2, *Factoryless manufacturing, global supply chains, and trade in intangibles and data* (pp. 81–120). Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Beissinger, Mark. (2021). *The urban revolution*. Book manuscript.

Beramendi, Pablo, Boix, Carles, & Stegmueller, Daniel. (2021). How Democracies Survive. Economic and Informational Foundations of Regime Stability. Working Paper.

Bernhofen, D. M., El-Sahli, Z., & Kneller, R. (2016). Estimating the effects of the container revolution on world trade. *Journal of International Economics 98* , 36–50.

Boix, Carles. (2003). *Democracy and redistribution*. Cambridge University Press.

Boix, Carles. (2011). Democracy, development, and the international system. *American Political Science Review 105* (4), 809–828.

Boix, Carles, Miller, Michael, and Rosato, Sebastian. (2013). A complete data set of political regimes, 1800–2007. *Comparative Political Studies 46* (12), 1523–1554.

Bourguignon, François, & Morrisson, Christian. (2002). Inequality among world citizens: 1820–1992. *American Economic Review 92* (4), 727–744.

Cortes, Guido M., Jaimovisch, Nir, & Siu, Henry E. (2016). Disappearing routine jobs: Who, how, and why? NBER Paper no. 22918. December. NBER.

Diamond, Larry. 2010. Liberation technology. *Journal of Democracy 21* (3): 69–83.

Dahl, R. A. (2008). *Polyarchy: Participation and opposition*. Yale University Press.

Elsby, Michael W. L., Hobijn, Bart, & Şahin, Ayşegül. (2013). The decline of the US labor share. *Brookings Papers on Economic Activity 2013* (2), 1–63.

Foa, R. S., & Mounk, Y. (2016). The danger of deconsolidation: The democratic disconnect. *Journal of Democracy 27* (3), 5–17.

Frey, Bruno S. (2010). *Happiness: A revolution in economics*. MIT Press.

Frey, Carl Benedikt, & Osborne, Michael A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114* (January), 254–280.

Geddes, Barbara. (2007). What causes democratization. In C. Boix and S. Stokes (Eds.), *The Oxford handbook of comparative politics* (Chapter 14, pp. 317–339). Oxford University Press.
Goos, M., Manning, A., & Salomons, A. (2014). Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review 104* (8), 2509–2526.
Guellec, Dominique, & Paunov, Carolina. (2017). Digital innovation and the distribution of income. National Bureau of Economic Research. Working paper no. 23987.
Gunitsky, Seva. (2015). Corrupting the cyber-commons: Social media as a tool of autocratic stability. *Perspectives on Politics 13* (1), 42–54.
Guriev, S., & Treisman, D. 2020. *Discourse dictators: How today's strongmen rule by distorting the news*. Book manuscript.
Hanson, G. H., Mataloni Jr, R. J., & Slaughter, M. J. (2005). Vertical production networks in multinational firms. *Review of Economics and Statistics 87* (4), 664–678.
Huang, Haifeng, Boranbay-Akan, Serra, & Huang, Ling. (2019). Media, protest diffusion, and authoritarian resilience. *Political Science Research and Methods 7* (1), 23–42.
Hummels, D. L., & Schaur, G. (2013). Time as a trade barrier. *American Economic Review 103* (7), 2935–2959.
Jungherr, A., Rivero, G., & Gayo-Avello, D. (2020). *Retooling politics: How digital media are shaping democracy*. Cambridge University Press.
Karabarbounis, L., and Neiman, B. (2014). The global decline of the labor share. *Quarterly Journal of Economics*, *129*(1), 61–103.
Katz, Lawrence F., & Margo, Robert A. (2014). Technical change and the relative demand for skilled labor: The United States in historical perspective." In Leah Platt Boustan, Carola Frydman, and Robert A. Margo (Eds.), *Human capital in history: The American record* (pp. 15–57). University of Chicago Press.
Kiatpongsan, Sorapop, & Norton, Michael I. (2014). How much (more) should CEOs make? A universal desire for more equal pay. *Perspectives on Psychological Science 9* (6), 587–593.
Kinkel, S., Lay, G., & Maloca, S. (2007). Development, motives and employment effects of manufacturing offshoring of German SMEs. *International Journal of Entrepreneurship and Small Business 4* (3), 256–276.
Kuran, Timur. (1991). Now out of never: The element of surprise in the East European revolution of 1989. *World Politics 44* (1), 7–48.
Kurer, T. (2020). The declining middle: Occupational change, social status, and the populist right. *Comparative Political Studies 53* (10–11), 1798–1835.
Layard, Richard, Mayraz, Guy, & Nickell, Stephen. (2008). The marginal utility of income. *Journal of Public Economics 92* (8–9), 1846–1857.
Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Broadway Books.
Londregan, John B., & Poole, Keith T. (1990). Poverty, the coup trap, and the seizure of executive power. *World Politics 42* (2), 151–183.
Lynch, Marc. (2011). After Egypt: The limits and promise of online challenges to the authoritarian Arab state. *Perspectives on Politics 9* (2), 301–310.
Malone, Thomas W., Laubacher, Robert J., & Johns, Tammy. (2011). The age of hyperspecialization. *Harvard Business Review 89* (7/8), 56–65.
Miguel, Edward, Satyanath, Shanker, & Sergenti, Ernest. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy 112* (4), 725–753.
Miller, M. K. (2012). Economic development, violent leader removal, and democratization. *American Journal of Political Science 56* (4), 1002–1020.

Norton, Michael. I., & Ariely, D. (2011). Building a better America—one wealth quintile at a time. *Perspectives on Psychological Science 6* (1), 9–12.

Nyst, Carly. (2018). Secrets and lies: The proliferation of state surveillance capabilities and the legislative secrecy which fortifies them—an activist's account. *State Crime Journal 7* (Spring), 8–23.

Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge University Press.

Przeworski, A., Alvarez, R. M., Alvarez, M. E., Cheibub, J. A., Limongi, F., & Neto, F. P. L. (2000). *Democracy and development: Political institutions and well-being in the world, 1950–1990* (No. 3). Cambridge University Press.

Qiang, Xiao. "The road to digital unfreedom: President Xi's surveillance state." *Journal of Democracy* 30, no. 1 (2019): 53–67.

Qin, Bei, Strömberg, David, & Wu, Yanhui. (2017). Why does China allow freer social media? Protests versus surveillance and propaganda. *Journal of Economic Perspectives 31* (1), 117–140.

Sage, D., and Diamond, P. (2017). "Europe's New Social Reality: The Case against Universal Basic Income." Policy Network Paper. February.

Siegel, David A. (2011). When does repression work? Collective action in social networks. *Journal of Politics 73* (4), 993–1010.

Starmans, Christina, Sheskin, Mark, & Bloom, Paul. (2017). Why people prefer unequal societies. *Nature Human Behaviour 1* , 82.

Svolik, M. W. (2019). Democracy as an equilibrium: Rational choice and formal political theory in democratization research. *Democratization 26* (1), 40–60.

Thewissen, S., & Rueda, D. (2019). Automation and the welfare state: Technological change as a determinant of redistribution preferences. *Comparative Political Studies 52* (2), 171–208.

Treisman, D. (2015). Income, democracy, and leader turnover. *American Journal of Political Science 59* (4), 927–942.

Tufekci, Zeynep. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Weingast, B. R. (1997). The political foundations of democracy and the rule of the law. *American Political Science Review 91* (2), 245–263.

World Bank. (2016). *World development report*. Digital Dividends.

Xu, Xu. (2020). To repress or to co-opt? Authoritarian control in the age of digital surveillance. *American Journal of Political Science 65* (2), 309–325.

CHAPTER 36

GOVERNING AI TO ADVANCE SHARED PROSPERITY

KATYA KLINOVA

Introduction

What governance levers can help to ensure that AI does not lower the demand for human labor, cutting off large swaths of the global population from their only source of income? This chapter examines the forces that currently give AI research and development (R&D) its labor-replacing focus. I analyze the monetary and non-monetary interests pursued by AI practitioners and the constraints their actions are subjected to, categorized into four modalities described by Lessig (1999): those arising from legislation, market conditions, social norms and "built architectures"—implicit and explicit codes by which the AI field operates. This chapter reviews the literature on the current state of each of these four sources of constraints as they apply to the field of AI, identifying gaps that stand in the way of operationalizing theoretical ideas about steering AI advancement towards inclusive economic outcomes. It argues that constructing worker-participatory ways to distinguish between economically sustainable and unsustainable AI investments is a pre-condition for enabling effective governance of AI in service of shared prosperity.

Why use governance to deliberately ensure AI does not diminish global labor demand? An alternative would be to let AI advancement stay on its current trajectory, which, as a growing number of leading economists and technologists agree, is likely to generate a large-scale redistribution of economic and political power towards a concentrated group of winners—a small handful of countries, firms, and individuals—eliminating jobs or lowering wages for an ever-growing share of the population (see, for example, Acemoglu & Restrepo, 2019, 2020; Korinek & Stiglitz, 2019; Altman, 2021). At present, a globally inclusive system for taxing and redistributing income and ownership of productive capacities does not exist. The political process necessary to enable the creation of such a system, if this process takes place at all, is very likely to lag behind the pace of technological change, resulting in large groups of the population getting left behind, which would undermine social stability and public trust in AI progress. Figuring out how to enable the beneficial advancement of AI while protecting and expanding access to good jobs is therefore a pressing

governance challenge. Rising to this challenge does not require imposing a ban on labor automation—the composition of available jobs can continue to evolve, with some jobs getting automated as long as new and better jobs replace them and are not associated with insurmountable skill barriers. But an abundance of well-paying, secure, and dignified jobs will be necessary at least until humankind has figured out how to robustly and on a global scale decouple the elimination of jobs from the elimination of dignity, status, and access to income.

Is it possible to steer AI away from labor displacement? The view presenting technological progress as an unalterable and unavoidable march of automation is common, but it overlooks a key consideration: the direction of technological change is a function of AI practitioners' choices made in pursuit of their interests, subject to constraints imposed—or not imposed—on them by the market, legislation, and the norms and structural features of the AI field. This chapter refers to "AI practitioners" broadly (including private and public sector researchers and engineers, entrepreneurs, venture capitalists, and corporations), and systematically examines the incentives and constraints that influence these actors' decisions around investing resources and time in development and deployment of labor-saving AI.

For examining the constraints AI practitioners are subjected to, the chapter uses Lawrence Lessig's pathetic dot theory as an organizing framework (Lessig, 1999). The theory proposes that all actions are constrained, or regulated, by four interacting forces: applicable legislation, the market, social norms, and "architecture." "Architecture" refers to "the way the world is" or the "built environment" where the activity of interest is performed—in the case of this chapter, the focus is on the "built-in" features of the AI development field. The four modalities of constraints interact and influence each other: the built environment can change as a result of legislation, and a change in legislation can be prompted by evolving social norms and expectations if those result in mounting political pressure. This chapter will specifically examine the influence of regulatory policy on market constraints and the impact of social norms and shared views in the AI field on defining its architecture and the ways things are done, highlighting the relevance of both of these types of interactions for the resulting direction of AI progress.

In Lessig's framework, legislation and the market are the first two forces that regulate any activity. AI advancement is no exception: the legislative and regulatory environment in which AI practitioners operate is determined by the applicable laws and policies which shape the market and influence the relative profitability of investing in various types of AI applications. Many policy decisions which superficially do not appear to direct AI R&D can either boost or reduce the incentives of AI developers to create labor-saving technology. For example, low interest rates make the investment in machines and software more appealing encouraging excessive automation (Stiglitz, 2014 as cited in Schindler et al., 2021), while tax regimes heavily favoring capital over labor can make it difficult to justify an investment in labor force expansion over automation even when producing the same output with machines is more expensive pre-tax (Acemoglu et al., 2020). Labor mobility restrictions limit labor supply and boost the incentives to develop labor-saving AI, generating little-studied cross-border spill overs of automation technology from countries with aging populations and restrictive immigration policies to countries struggling to produce sufficient numbers of formal sector jobs for their young and growing labor forces (Pritchett, 2020). But successfully re-orienting policies to steer AI advancement towards an economically inclusive trajectory requires not only a recognition of the labor demand-depressing

side effects of current policies but also a construction of a better understanding of how to practically distinguish labor-friendly AI from inequality-producing AI.

The project of governing the direction of AI in service of shared prosperity also needs to pay attention to something more subtle and much less well described than regulatory impacts on automation incentives. The chapter argues that the structural features of the AI field, which include what Lessig defines as norms and architecture, bias its focus towards labor-saving applications. AI practitioners often split their time between the public and private sector, but even those who work only in the private sector are driven not only by the profit motive. They are also influenced by the field's features—a web of its norms, shared aspirations, commonly accepted ways to measure progress, and ideas about desirable future for the human society. "The field's goal has always been to create human-level or superhuman AI," states Stuart Russell, one of the preeminent AI scientists, in the opening of his latest book (Russell, 2019). The structure of the AI field orients towards this explicit goal of matching and exceeding human performance on all basic capabilities. In addition, tech elites' dominant political view favors redistribution, which is seen as a solution to an automation-induced loss of labor income, but strongly disfavors government regulation of technology (Broockman et al., 2019). This conjunction of the fields' goals and political beliefs fuels the labor-saving bias of AI advancement in a multitude of direct and indirect ways.

The rest of this chapter proceeds as follows. First, after reviewing the economic literature raising concerns over AI's current direction, it describes the practical difficulties around differentiating labor-displacing and labor-complementing technologies for legislative or market-regulating purposes, exacerbated by the absence of empowered participation by workers in the AI development process. Next, the chapter reviews policies channeling AI development towards excessive labor substitution. Lastly, the chapter examines the stated objectives of the leading AI labs and attempts to identify which socio-technical factors (or, using a term from Jasanoff & Kim, 2015, "imaginaries") might have shaped those objectives, and how they can be altered by governance mechanisms in service of re-orienting the field towards producing shared prosperity-compatible outputs.

Policy, Market Incentives, and the Difficulty of Defining Labor-friendly AI

Practical ambiguity of the "human-augmenting AI" concept

The possibility of using AI for automating human labor does not make AI advancement non-labor-friendly by definition. Automation is a phenomenon that long predates AI and that has enabled a broad-based rise in productivity and living standards. What is different about the wave of labor automation enabled by AI is not only that AI allows for a dramatic expansion in automation possibilities, potentially leading to a meaningful acceleration in the pace of automation, but also that AI advancement is happening against the backdrop of rising inequality and worsening labor market outcomes for an ever-increasing share of the population—trends that AI, on its current trajectory, is poised to exacerbate.

In the last four decades, growth in the advanced economies has not been inclusive: labor shares of national incomes have steadily declined, wages of workers without college degrees have stagnated or even declined in real terms, the number of middle-paying jobs has decreased, giving rise to the gap between average and median worker compensation (Autor et al., 2020). Acemoglu and Restrepo (2021) document robust evidence suggesting that at least in the case of the U.S., the majority of the change in the economy's wage structure is explained by the relative wage declines of workers in industries that experienced rapid automation. In their earlier work, same authors show that unlike in the four decades following the Second World War, when the volume of waged tasks automated every year was matched by newly created waged tasks for humans, in the last three decades in the U.S., the pace of automation has measurably accelerated and consistently eclipsed the pace with which new waged tasks appeared. This has contributed to a decline in labor demand and compensation for workers in automation-exposed occupations (Acemoglu & Restrepo, 2019).

AI-induced technological change, like any other type of technological change, can be expected to bring about two effects: an increase in the overall economic output and a redistribution of income between factors of production, where factors of production usually include labor, capital, and sometimes land. Labor can be further disaggregated into groups of workers, for example, by skill level. If labor receives at least some of the gains from technologically induced output increase, technological change is referred to as labor-using, and in the opposite scenario as labor-saving. For example, technological change that powers labor-displacing automation but does not create a compensatory labor demand would be labor-saving. Following Korinek et al. (2021), this chapter uses the terms "direction," "focus," and "orientation" of the AI field as a reference to whether the technological change it generates is labor-saving or labor-using.

Further, if labor benefits from technological change relatively more than the other factors of production, technological change is deemed biased in favor of labor. Of course, not all technological change is biased in favor of labor—it can often be biased in favor of capital, or only certain kinds of labor. Technological change can complement certain skill groups and raise their gains while reducing the demand for other skill groups, in which case it is deemed skill-biased. For example, the proliferation of personal computers and the development of word processing and planning software tools complemented the skills of people in managerial and other occupations but reduced the demand for typists and secretaries. Skill-biased change is likely to deepen economic inequality.

While technological change is not the only factor that influenced real wage stagnation and wage declines for workers with lower levels of formal educational attainment in recent decades, the persistence of this trend coupled with accelerating automation of tasks associated with middle-playing jobs suggest that in its latest form technological change has been biased against formally lower-skilled groups. To ensure that this inequality-deepening trend is not accelerated by the advancement of AI, leading economists have been calling to create AI technologies that are labor-complementing and not labor-saving because the latter kind would decrease the demand for human labor, leading to a reduction in either employment, wage levels, quality of jobs, or some combination of all of the above (Korinek & Stiglitz, 2019; Acemoglu & Restrepo, 2020).

An increase in the demand for human labor is generally associated with an increase in employment, wages and improved working conditions, so technologies that lead to an increase in labor demand would benefit labor, at least in absolute terms, if not relative to

other factors of production. But practically executing on the recommendation to develop labor demand-boosting AI is difficult because the effect of new technology on labor demand is highly uncertain ex-ante due to the presence of multiple second-order effects, possible variation in deployment contexts, a variety of ways basic research can be used in final applications, and unknowable counter-factual scenarios of technological advancement.

Let us examine each of these sources of uncertainty in turn. Second-order effects of technological change refer to its indirect impacts: outside of jobs directly created or cut by a company that develops or introduces a new technology, often there are jobs created or cut (or wages increased or reduced) elsewhere in the economy, for example, by the company's suppliers, clients, or competitors (see a detailed categorization of indirect effects in Klinova & Korinek, 2021). If the company's introduction of new technology creates productivity gains and those are passed on to consumers in some form (e.g., as reduced prices, improved quality, or new products), that can free up consumers' income to be spent on other goods in the economy, creating new labor demand in corresponding sectors. Productivity gain does not always get fully passed on to consumers, especially in monopolistic markets in which leading AI companies tend to operate. Acemoglu and Restrepo (2019) also warn of the recent proliferation of what they refer to as "so-so" technologies—those that displace human labor but do not create a meaningful productivity gain, thus failing to give rise to a compensatory labor demand elsewhere in the economy.

Early-stage AI research can enable a wide range of applications down the road, including both those biased in favor of and against labor. This adds to the difficulty of practically steering AI advancement in service of increasing employability of economically vulnerable workers. Moreover, the same high-level application can be used to replace or augment human labor depending on the deployment context. Lastly, evaluating a tentative impact of new technology on labor demand based on a comparison with the status quo can be misleading because such comparison does not take into account the impact of technologies the development of which would happen thanks to the creation of the new technology in question, nor the impact of alternative technologies that could be developed by people who are busy with the creation of the technology in question.

Despite the presence of all these uncertainties about the impact of AI applications on labor demand, AI companies have begun to pick up on the growing concern around labor-saving AI and increasingly describe their products as labor-augmenting. The meaning ascribed to the term is frequently vague, suggestive of AI that "helps" or "assists" workers by, for example, making them more productive or reducing the number of workplace accidents. This broadly matches the economic definition of labor-augmenting technologies as technologies that increase the marginal product of labor. However, "worker-augmenting" AI does not guarantee that workers would receive any gains as a result of its introduction and does not ensure that workers would not be made worse off. A prime example of technologies that are often positioned as worker assisting, productivity increasing and safety improving are workplace surveillance solutions, usually described by their producers as tools for "worker productivity monitoring."

As documented by Nguyen (2021), workplace monitoring technologies enable work speed-ups and intensification, the creation of excessively punitive work environments and the shifting of risks from employer to employee. For example, flexible algorithmic scheduling allows employers to offload the risk of a slow day onto workers whose shifts are scheduled or cancelled at the last minute based on fluctuating customer demand projections. In

the context of the workplace principal–agent problem (where the employer is the principal and the worker is the agent), the principal's imperfect ability to observe the agent's effort is a source of power for the agent that otherwise typically holds very few powers, especially in a non-unionized setting. Imperfect ability to observe an agent's effort gives the employer a motivation to incentivize high performance by treating the agent to dignified working conditions, performance bonuses, etc. When an agent's effort is perfectly observable, the principal has less of an incentive to reward high performance; she can set targets for the display of effort and penalize workers not meeting them (Gerety, 2020). In other words, the use of "sticks" in a workplace under total surveillance is likely to dominate the use of "carrots."

Workers being monitored by assistive AI might or might not be aware that they are training an algorithm to become a better substitute for them. Even when workers are fully aware of the training they are participating in, like, for example, human drivers helping self-driving cars navigate difficult and ambiguous situations, they are nearly never recognized as co-creators of the resulting technology (Lanier, 2014). They do not hold IP rights for their know-how, do not receive royalties every time the data they generate is used, are generally not granted equity shares, and are not allotted the praise and social status of AI practitioners.

To avoid a proliferation of technologies that claim to augment workers but in practice enable employer overreach and exploitation and strip workers of privacy and power in the workplace, it is necessary to introduce measurement and disclosure requirements around AI systems' impact on labor. Aside from an analysis of the magnitude of second-order impacts on labor demand referenced above, the disclosures should be required to contain results of independently carried out surveys of workers' experience with the AI system in question, their involvement in decisions around the introduction of AI into the workplace, as well as channels of recourse and contestation available to them. This is necessary to ensure an empowered participation of workers in the design and deployment of AI systems that are poised to affect them, as well as because affected workers are likely to have first-hand insight into whether an AI system in question benefits or hurts them.

In game theoretical terms, today the claim that "our AI augments humans" that is made by a growing number of AI companies amounts only to a cheap signal, and hence it cannot be used by regulators or potential buyers to differentiate AI systems that expand access to good jobs from tools that benefit capital owners but disadvantage workers. Substantiating the claim with tangible and transparently reported metrics around the impact of an organization's AI development and deployment efforts on labor demand and job quality would allow it to meaningfully differentiate itself from others with a credible signal. Organizations genuinely committed to a mission of augmenting and complementing human workers should want their commitment to be seen as credible and thus should be willing to report on their impact on the availability and distribution of good jobs in the economy.

There are efforts currently underway to develop robust ways to measure labor demand impact of a given AI-developing organization (see, for example, Partnership on AI, 2021). Such measurement approaches can be used by AI practitioners in the private and public sectors to inform their assessment of the economic consequences and social sustainability of new products and services and hopefully to guide their product choices in a prosocial direction. Governments can use such measurement frameworks to inform their R&D investments and industrial policy, as well as introduce a rule to procure AI only from companies that assess and disclose their labor market impact (Seamans, 2021). Importantly,

governments should also pay attention to how their policies across the board might be incentivizing the development of labor-saving technologies. The next section turns to the interaction between Lessig's first two forces—legislation and markets—to give an overview of how the market incentives faced by the AI industry are shaped by policy choices, biasing the advancement of AI against labor.

Policies boosting market incentives for the development of labor-saving AI

Economic incentives that commercial AI development, including the development of labor-saving AI, is faced with are determined by multiple factors, including the present and expected conditions of the global economy, state of competition and regulation, the demographic situation, and more. Consequently, policies of special relevance to shaping the direction of AI development include tax policy, R&D and industrial policy, interest rate policy, government procurement practices, policies around migration, and what Korinek et al. (2021) refer to as "rules of the game," or policies that affect the returns on factors of production, such as labor legislation, competition laws, rules regulating corporate governance, etc. All of these can raise or lower the relative commercial appeal of investing in labor-saving technologies. For example, low interest rates stimulate capital investment into technology and equipment and large government orders for machines that substitute for human workers can accelerate the development of such machines, etc. Policies introducing distortions deserve special attention because they can lead to excessively high (or low, depending on the direction of the distortion) levels of automation. This subsection will focus on two policies giving rise to the biggest distortions in the pace of automation: policies around taxation and labor mobility.

In much of the industrialized world labor is taxed more heavily than capital, which makes replacing a higher portion of the workforce with machines financially appealing for businesses, provided that the net present value (NPV) of technology development costs does not exceed the NPV of the tax payments that will be saved. That said, if taxes are set optimally (which for many countries might mean capital is taxed at a lower rate than labor), the level of automation will also be optimal in absence of other distortions. Acemoglu et al. (2020) show that, at least in the U.S.—a country with outsized influence on AI R&D—the effective tax rate on labor was too high in the 2010s, while the effective tax rate on capital was too low, leading to excessive levels of automation compared to the socially optimal level. They note that even if tax rates were set optimally going forward, it would still be welfare-improving to reduce the resulting equilibrium level of automation because the starting point would be one with already excessive levels of labor-saving technologies brought about by tax distortions present throughout the 2010s.

Because the United States plays a dominant role in AI development and because of low marginal costs of deploying AI applications around the world once they have been developed in the U.S., the country's distorted tax code is effectively "exported" to the rest of the world through excessive levels of automation spilling over to low- and middle-income countries struggling to create a sufficient supply of formal sector jobs for their young and growing labor forces. For example, recent evidence analyzed by Diao et al. (2021) suggests

that global technological trends exported to African countries induce local firms to employ capital-intensive technologies, which are inappropriate in the context of those countries' comparative advantage or workforce needs.

The borderless nature of technology deployments prompts a discussion of potential policy-induced incentive distortions not only from the point of view of a single country but also of the entire planet. Pritchett (2020) points out what he describes as "the biggest price distortion ever": the disparity between wages in high- and low-income countries. Clemens et al. (2019) show that labor price differentials between rich and poor countries exceed any current or historic trade tariffs or carbon price distortions by orders of magnitude. This leads to a situation where private sector actors producing technologies in high-income countries face a distorted labor supply curve: the labor supply they effectively respond to does not reflect the global supply of labor, while the technologies they produce do spread globally relatively quickly. As theoretically shown by Acemoglu (2010), labor scarcity encourages strongly labor-saving technological change, which can manifest in lower employment, wages or under-employment domestically. If labor-saving technologies spread across the globe—which they often do—that generates a negative externality for countries struggling to create an adequate provision of good jobs, especially for their youth.

Pritchett (2020) also points out that, aside from giving rise to an excessive spread of labor-replacing technologies, distorted prices of labor can and do lead to the creation of labor-shifting technologies. For example, self-service kiosks widely deployed in restaurants, grocery stores, and airports around the world do not exactly automate the cashier's or check-in agent's jobs; instead, they shift almost entirely the same set of tasks onto the customer, who does not get paid for executing them, and might or might not get any amount of compensating benefit in the form of lower wait time or lower price. In other cases, technology might be partially shifting paid work into unpaid work. For example, calling customer service is increasingly associated with navigating automated self-service voice menus, which can cost the customer a lot more time and frustration with little benefit. Work-shifting applications often fall into the "so-so technologies" category (a term from Acemoglu & Restrepo, 2019): they reduce paid employment but do not compensate for that with any meaningful productivity boost to the economy. Task-based platforms for ride-sharing, delivery, home repairs, or online tasks like data labeling also offer a wealth of examples of jobs being broken down into compensated and uncompensated parts. Workers on those platforms are not paid for their time spent waiting or searching for the next task, for learning how to complete the task, for maintaining and repairing their equipment, let alone for taking a lunch break or a sick day (Gray & Suri, 2019), while all of those activities would be compensated in case of a standard employment contract.

The above discussion suggests that more immigration would benefit both high- and lower-income nations: restricting immigration is a weak strategy for protecting jobs at home. Large and growing wage disparities across the countries incentivize offshoring of jobs to lower-wage countries, shifting of jobs to unpaid work and excessive creation of labor-saving technology that undermines wages and quality of jobs at home and spreads beyond a single country's borders. For example, the U.S. could bring in 160,000 immigrants to close the truck drivers shortage projected to accumulate by 2028 by the American Trucking Association (Costello & Karickhoff, 2019), or it can develop autonomous driving technologies and displace all 3.6 million people employed as truck drivers in the U.S., and likely many more millions employed in this job abroad. And while autonomous driving

might come with important benefits, such as improved road safety that might make the technology desirable despite the lost jobs, similar dynamics are at play creating distorted incentives for the development and deployment of worker-replacing technologies with much more ambiguous benefits, for example, those replacing nurses (Mani et al., 2021), even though the shortage of care workers could be beneficially reduced by expanding cross-border labor mobility.

Many economists agree with the hypothesis that private sector-driven technological change, responding to societal scarcities as reflected by (non-distorted) market prices, can and has generated incredible innovation which supported the rise in living standards observed over the last two centuries. However, there is no theoretical basis for stating that the market process generates an optimal trajectory for technological change in the long term (Korinek, 2019). And if the price signals are distorted not in favor of labor, which they presently are, as has been discussed in this section, the market is likely to deliver excessive automation beyond socially optimal levels, directing AI to economize on false scarcities, eliminating good jobs at a time when much of the world is struggling to generate enough of them. But market incentives distorted by misguided policies are not the only factor fueling the AI field's dangerous focus on churning out labor-saving and worker-exploiting technologies. The next section will examine the contribution of the other two forces on Lessig's list—social norms and structural features of the environment in which AI development happens.

AI Field's Orientation, Norms and Structural Features that Reinforce It

This section reviews how social norms and architecture of the AI field influence the direction of AI development. AI field's norms arise from a shared understanding of what problems are worth tackling and can bring praise and recognition, explicit and implicit definitions of success and state-of-the-art performance, which this section examines. Architecture, in Lessig's definition, is "the way the world is," the "built environment" in which AI development and deployment happens and its structural features. The architecture of the environment and its norms do interact with each other, just like the legislative and market constraints which were discussed in the previous section, but in the context of the AI field, the norms-architecture interaction is notably tight. Lessig (1999) suggested a useful way to differentiate between them: norms constrain or direct the agent's actions only if the agent is aware of them, while architecture's impacts are not predicated on awareness. For example, in the context of AI, a rare practitioner is not aware of Moore's law, but its powerful impact would be experienced by AI practitioners with or without that awareness, making it a feature of the "built environment." Similarly, the present homogeneity of the AI field in terms of demographic and socio-economic backgrounds impacts what kind of problems get taken up by the field's actors and what kind of consequences of AI deployments get examined or ignored—that impact takes place whether or not the actors in the field are aware of its homogeneity. Unlike that, an AI entrepreneur who is unaware of the venture capital community's appreciation of exponential user growth is less likely to

emphasize it in her business pitch over profit generation, placing this case in the category of adhering to norms and commonly held expectations.

Governing ideas and norms of the AI field

Definitions of success and choice of problems to tackle

A field's orientation is influenced by which problems are selected to be tackled by its members and how they define progress. Whether the field takes up addressing social or commercial challenges, and what kind of achievements are associated with reputational gains and prestige, influence the impacts the field generates on the society, including its economy and labor market conditions.

In his 1988 Presidential Address to the Association for the Advancement of Artificial Intelligence, Turing Award winner Raj Reddy recounted the following story: "In 1966, when I was at the Stanford AI labs, there used to be a young Ph.D. sitting in front of a graphics display working with Feigenbaum, Lederberg, and members from the Chemistry department attempting to discover molecular structures from mass spectral data. I used to wonder, what on earth are these people doing? Why not work on *a real AI problem* like chess or language or robotics? What does chemistry have to do with AI?" (Reddy, 1988, emphasis mine). "Now we know better," Reddy continued, before proceeding with a long list of notable applications of AI in various fields. That list has grown tremendously since 1988, but the statement remains a testament to the existence of shared notions or norms around what problems are "real" or worth pursuing.

Which problems are considered worthwhile in today's field of AI? One way to ascertain that is by reviewing how the leading AI labs describe their work. Table 37.1 contains self-descriptions of private and academic labs that were among the top 10 places of affiliation of the authors of papers accepted to the 2020 International Conference on Machine Learning, one of the field's primary conferences (Ivanov, 2020).

Among the university-based AI labs mentioned in Table 36.1, two—MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and Carnegie Mellon University (CMU)—describe themselves with language that could be interpreted as expressing an intent to assist or complement humans, by mentioning building "things that help humans" and pioneering "research in computing that improves the way people work, play, and learn" respectively. Two other academic labs—Princeton Visual AI lab and Stanford AI lab—list goals around improving human–AI collaboration. Among the leading private AI labs, only one (Microsoft Research) declares an explicit intent to strive for human complementarity, describing itself as "[p]ursuing computing advances to create intelligent machines that complement human reasoning to augment and enrich our experience and competencies." While that description currently stands in stark contrast with those that explicitly or implicitly declare goals around automating human activities starting with basic abilities, human augmentation language alone is too vague to serve as a strong signal of, or commitment to, a qualitatively different path of AI development, as was discussed previously.

A few of the self-descriptions listed in Table 36.1 explicitly mention advancing the state-of-the-art in the field of AI. What is considered to be state-of-the-art performance in AI? The "Technical Performance" chapter of the most recent AI Index report (2021) gives a

Table 36.1 Self-descriptions of the top 10 AI labs by a number of affiliated authors whose papers were accepted to ICML 2020 as identified by Ivanov (2020).

Organization	Self-description
Google	[W]e're conducting research that advances the state-of-the-art in the field, applying AI to products and to new domains, and developing tools to ensure that everyone can access AI (https://ai.google/about/).
MIT	MIT's Computer Science and Artificial Intelligence Laboratory pioneers research in computing that improves the way people work, play, and learn (https://www.csail.mit.edu/).
Stanford University	[Stanford Artificial Intelligence Lab] promotes new discoveries and explores new ways to enhance human-robot interactions through AI (https://ai.stanford.edu/about/).
UC Berkeley	The Berkeley Artificial Intelligence Research (BAIR) Lab brings together UC Berkeley researchers across the areas of computer vision, machine learning, natural language processing, planning, control, and robotics (https://bair.berkeley.edu/).
DeepMind	We're a team of scientists, engineers, machine learning experts and more, working together to advance the state of the art in AI (https://deepmind.com/about).
Microsoft	Pursuing computing advances to create intelligent machines that complement human reasoning to augment and enrich our experience and competencies (https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/).
Carnegie Mellon University	CMU has spent decades building a culture where people care about using technology to solve real problems. More than half a century ago, Allen Newell and Herb Simon had a vision for a general problem-solver for the human race. Since then, their vision has become a filter: people attracted to building solutions to real-world problems come here. The result? One of the world's largest collections of people determined to build things that help humans (https://ai.cs.cmu.edu/about).
Princeton University	We work on developing artificially intelligent systems that are able to reason about the visual world. Our research brings together the fields of computer vision, machine learning, human-computer interaction as well as fairness, accountability and transparency (https://visualai.princeton.edu/).
Facebook	We're advancing the state-of-the-art in artificial intelligence through fundamental and applied research in open collaboration with the community (https://ai.facebook.com/research).
University of California Los Angeles	The StarAI lab [Statistical and Relational Artificial Intelligence Lab] performs research on Machine Learning (Statistical Relational Learning, Tractable Learning), Knowledge Representation and Reasoning (Graphical Models, Lifted Probabilistic Inference, Knowledge Compilation), Applications of Probabilistic Reasoning and Learning (Probabilistic Programming, Probabilistic Databases), and Artificial Intelligence in general (http://starai.cs.ucla.edu/).

Source: Collected by the author from organizations' websites (accessed on June 25, 2021).

detailed picture: it describes the progress in today's main subfields of AI, namely computer vision, language, speech, concept learning, and theorem proving. In other words, advancing the state-of-the-art in AI means improving machines' ability to imitate basic human capabilities: to see, speak, hear, write, and reason. In these subfields, performance benchmarks are designed to enable an explicit comparison with human performance. The following subsection gives an overview of key benchmarks.

Orienting benchmarks of the AI field

Benchmark datasets specify the goals the AI research and development community optimizes their work for. Large high-quality datasets are difficult and costly to compile anew, which ensures the enduring popularity of existing publicly available ones. Their use is further incentivized by associated prizes and challenges, some of which happen annually and award not only sizable monetary amounts but the status of the field's leader. Chasing state-of-the-art performance on benchmark datasets has emerged as a common goal in the subfields that constitute today's AI R&D (Raji et al., 2020). Examining key benchmarks is therefore instructive for understanding what the AI field is presently building towards and what it defines as success.

This subsection reviews benchmarks used to assess progress in today's key subfields of AI: vision, speech, language understanding, and reasoning. The benchmarks listed below—ImageNet, SuperGLUE, LibriSpeech and VCR—are among those that the AI Index report (2021) references to describe the state of AI's technical performance.

ImageNet is an image database of over 14 million labeled images, which, according to the ImageNet website, contains "hundreds and thousands" of image examples for each noun in the English language. The dataset is available for free for non-commercial use and is used to train object detection and image classification algorithms. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates the performance of such algorithms and measures the progress of computer vision as a field (Russakovsky et al., 2015). Two metrics are most commonly reported on: Top-1 accuracy, or the percentage of times when an image label predicted by an algorithm matches the correct label indicated in ImageNet, and Top 5 accuracy, or the percentage of times a correct label is found among algorithm's five best predictions. Top 1 accuracy has grown from just over 60 percent in 2013 to 86.5 percent in 2021; Top 5 accuracy has grown from just under 85 percent in 2013 to 97.9 percent in 2021, "beating" human performance.

SuperGLUE is a benchmark for evaluating the progress of AI systems in English language understanding. It was introduced in 2019 as a more challenging version of the GLUE benchmark after the performance on GLUE had surpassed the level of non-expert humans. SuperGLUE assigns a single score summarizing the progress in eight language-understanding tasks, each of which comes with publicly available standardized training, development, and test datasets (Wang et al., 2019). Best performing models are listed on the SuperGLUE leaderboard. At the time of writing, the leaderboard is topped by a team with a score of 90.4, which is just above SuperGLUE's "human baseline" of 89.8.

LibriSpeech is a dataset consisting of over a thousand hours of public domain audiobooks and is used to train and evaluate speech recognition systems (Panayotov, 2015). Similarly to the benchmarks discussed above, LibriSpeech has a corresponding leaderboard that ranks models by how well they perform on the LibriSpeech test set. Models with the lowest

transcribed word error rate at the top. As of April 2, 2021, the best-performing models get 1.4 percent of words wrong in a controlled environment and 2.6 percent wrong in a noisy environment.

Visual Commonsense Reasoning (VCR) is a dataset consisting of 290,000 question–answer problems derived from 100,000 movie scenes. It is used to evaluate how well machine learning (ML) models can imitate a basic human ability to discern the context and situation from an image. For each test image and a corresponding question about what is happening in the image, an algorithm is expected to choose a correct answer from a selection of four and pick a correct rationale from a selection of four. For example, a question about an image with two people and a basket of money in front of them is "How did person 2 get the money that's in front of her?" The correct answer is: "Person 2 earned this money playing music," and the correct rationale is: "Person 1 and Person 2 are both holding musical instruments and were probably busking for that money" (Zellers, 2019). According to the VCR Leaderboard, no ML model has so far crossed the "human performance" benchmark at 85 points. That said, improvement in model performance has been impressive: the best model's score has gone from 44 in 2018 to 70.8 in 2020 (AI Index, 2021).

All the benchmarks described above unambiguously map to the goals of automating or imitating basic human abilities. Corresponding leaderboards explicitly recognize which AI models managed to reach or cross a "human baseline." This normalizes and incentivizes a "competition" between humans and AI models in which the world's best AI researchers chase the goal of creating systems that excel in tasks which most of the 7.8 billion people living today can perform and use every day to earn a living. Of course, one could argue that it is possible to construct an AI system that leverages computer vision, speech recognition, and other technologies to complement human workers, raise their productivity and make them more valuable to the labor market, and not to displace them from jobs. But that will not be achieved without the field exchanging its current goals—focused on beating people at their basic abilities—for goals aimed at increasing peoples' productivity (Siddarth et al., 2021). I will turn next to a discussion of why goals in the AI field are being set the way they are, and what alternative benchmarks have been suggested.

The tradition of evaluating progress in AI in terms of matching and "beating" human performance on basic tasks is a long-standing one: in his 1988 Presidential Address to AAAI, Raj Reddy notes that at any given point, the AI field's accomplishments can be measured by "assessing the capabilities of computers over the same range of tasks over which humans exhibit intelligence." Mani et al. (2021), reviewing the substantial progress made by the field of AI since 1988 on the grand challenges posed by Raj Reddy in his presidential address (such as building a world champion chess machine and an accident-avoiding car), provide examples of different goals the AI field could set for itself. For instance, Frank Chen proposes to shift the focus of the goal of AI research away from "surpassing" humans and towards "human + AI = better together" vision, combining the strengths of machines and humans. He suggests judging the progress based on the ability of human+AI teams to perform better than either humans or machines on their own. In the same volume, Steve Cross proposes a "Reddy test" for teams, judging progress based on appropriate high performing team criteria of any given domain—for example, a human+AI team of pilots and air traffic personnel being able to successfully handle an unprecedented situation (Mani et al., 2021).

Adoption of benchmarks and goals around human+AI teams and their elevation into a category of coveted and prestigious goals to chase by the field can be meaningfully helped

by structuring government-funded challenges around them because, as noted by Mani et al. (2021), grand challenges act as important "compasses" for AI researchers, especially young researchers looking for worthwhile problems to work on. Government R&D challenges are known for having been able to kick-start entire fields in AI and prompt significant private sector investment, amounting in some cases to hundreds of billions of U.S. dollars. One of the most famous ones is the DARPA Grand Challenge that first ran in 2004 "with the goal of spurring on American ingenuity to accelerate the development of autonomous vehicle technologies that could be applied to military requirements" (DARPA, 2005). In a little over a decade since the first challenge, autonomous driving has attracted tens of billions of U.S. dollars in private investment and counting (Kerry & Karsten, 2017), while DARPA continued to pursue the Grand Challenges model to spur on R&D in areas such as robotics and fully automated network defense (DARPA, 2014). A well-designed challenge around labor-enhancing AI could similarly serve as a powerful catalyzer for the advancement of worker-empowering AI applications. As applicable, challenges can come with benchmark datasets deliberately designed to evaluate the ability of an ML model to assist a human worker and boost her productivity in non-exploitative ways, instead of trying to outperform humans on their basic abilities.

Sociotechnical imaginaries driving AI development

Aside from grand challenges and chasing state-of-the-art performance targets, influential ideas about the desirable technological future for society also serve as a powerful orienting force for AI practitioners. Those ideas are heavily shaped by the technologists themselves, as well as by cultural artifacts, such as hit science fiction stories. Notable among these stories were *Star Trek*, a cult film franchise and television series that aired in the U.S. from the mid-1960s through the mid-2000s, and Isaac Asimov's *Robot* stories, which Star Trek is partially based on. Asimov was the first science fiction author to present robots favorably; in his stories engineers are enlightened characters. This vision inspired many to pursue robotics. Marvin Minsky, a Turing Award recipient who is widely regarded as a key figure in the history of AI, is said to have decided to become a computer scientist and work on robotics after reading Asimov's science fiction stories (Saadia, 2016). To this day, start-ups successfully raise capital to build devices "from *Star Trek*" (see, for example, Sweeney, 2019), and even executives of the largest companies openly declare replicating *Star Trek* technology as their orienting vision—for example, Amit Singhal, Google's former executive in charge of Google Search from 2000 till 2016, wanted Google Search to work like his "dream *Star Trek* computer" (Luckerson, 2016).

The fact that the public relations department of a major publicly traded corporation thought it desirable to use a childhood dream of a technologist in charge of their core product as a focal point of this product's public positioning suggests that contemporary society commonly views the intentions to build *Star Trek* technology—as well as people and organizations declaring such intentions—very favorably. In the 2015 book, *Dreamscapes of Modernity*, edited by Jasanoff and Kim, desirable visions of the future that do not simply represent individual vanguard visions but become collective reference points and anchors for future projects are referred to as "sociotechnical imaginaries." Jasanoff noted that imaginaries are "performed" collectively and that they co-produce the reality of not only the world we live in today but also of the "known, the made, the remembered and the

desired worlds." The co-production entails that sociotechnical imaginaries are simultaneously "instruments and products" of a collective understanding of what the world and society should look like (Jasanoff & Kim, 2015). *Star Trek* has for decades been serving as a collective reference point used by Silicon Valley engineers, entrepreneurs, and venture capitalists (VCs). If it hints at a contour of an imaginary performed by the U.S. tech industry professionals—a community with outsized influence on the direction of AI—it is suggestive not only of what kind of technological tools they dream of building but what kind of social order they aspire to on behalf of human society.

In *Star Trek*'s view of the future, members of the United Federation of Planets—founded on the principles of peace, justice, and equality—live in a post-scarcity world, liberated from the necessity to work. They possess a "Replicator"—a device capable of producing almost any desired good instantly at no cost. *Star Trek* was one of the very first stories in science fiction that painted a picture of a positive, utopian future in which technological progress and automation of human labor were beneficial underpinning forces. But, as pointed out by Saadia (2016), the real miracle of *Star Trek* was not its technology, but the social order that enabled everyone to share in the fruits of the progress.

Notably, the *Star Trek* series contains very little on what policy choices enable and sustain the social order the Federation enjoys. We obviously do not know if in real life the automation of human labor would ever lead to an egalitarian social order where everyone is entitled to share in the abundance. If history is any indication, the likelihood of economic and political power getting shared through voluntary redistribution from winners to losers, unprompted by long and difficult power struggles with highly uncertain outcomes, is very low. And yet, the technology industry leaders, at least in the U.S., seem to place a lot of hope on monetary redistribution, while feeling deeply opposed to measures and institutions redistributing political power, which is necessary for the stability and continuity of not only the monetary redistribution schemes but of democratic governance as well.

Brookman et al. (2019) surveyed 600 U.S. technology companies' founders and executives, most of them millionaires, who have raised more than $19.6 billion in venture capital investment. The majority of the study participants (62.1 percent) chose "Don't Regulate and Do Redistribute" as the best description of their views: they do expect the government to tax and redistribute wealth and they are supportive of universal healthcare and programs benefiting the poor, but they do not feel favorably about government regulation, seeing it as doing more harm than good especially when it comes to regulating technology product markets and the labor market. An overwhelming majority of respondents would like to see the strength of unions decrease and consider it to be at present excessively hard to fire workers.

Whether the "Don't Regulate and Do Redistribute" worldview subscribed to by the outsized share of U.S. technology leaders is based on a sincere belief that unregulated technological advancement and monetary redistribution are sufficient for attaining an equitable distribution of power in society, or it is simply a posture deliberately adopted for self-serving reasons, this worldview is likely to be material for any effort of governing AI in service of shared prosperity. There is much more to be understood though about the "mechanics" of how it influences the direction of AI. As Jasanoff and Kim wrote in *Dreamscapes of Modernity* (2015), "particularly empty of theoretical guidelines is the domain that connects creativity and innovation in science, and even more technology, with the production of power, social order, and communal sense of justice." The development of

theoretical guidelines and frameworks that illuminate this crucial connection is important for society's ability to take control of governing AI in a democratic way.

Structural features of the AI field

What structural features of the AI field prompt and sustain its focus on imitating basic human abilities? To achieve recognition in the field that celebrates state-of-the-art performance, one needs to demonstrate that performance by running a high number of experiments, which is one of the features papers accepted to top conferences in AI are characterized by. This puts academic researchers at a disadvantage compared to their industry peers, because the former group, with rare exceptions, is much more restricted in its access to computing power. This is a major factor contributing to the present state of the field in which the industry, and not academia, is central to AI research activity (Reich, 2021). Between 2010 and 2019, the share of graduating PhDs in AI in the U.S. and Canada going to the industry grew from 44.4 percent to 65 percent (AI Index, 2021, cited by Reich, 2021). This "brain drain" away from the academic AI research centers can have far-reaching implications for what real-world problems get tackled by AI researchers and what problems remain without sufficient attention.

The significance of restrictions placed on AI practitioners working in the for-profit sector is highlighted by Rakova et al. (2021). The study presents the results of 26 semi-structured interviews with the AI industry practitioners who advance responsible AI practices in their organizations either as a part of their formal job description or voluntarily. They report having to distill what they do into standard metrics used by the industry (e.g., number of clicks, acquired users, churn rate) as one of the key barriers to achieving progress in their work. They are commonly asked to measure the impact of the responsible AI efforts in terms of revenue generated. Product teams they are frequently embedded into are usually pressured to deliver within fast-paced development cycles that incentivize the use of success criteria that are easier to measure and discourage paying attention to long-term societal outcomes. Industry organizations surveyed by Rakova et al. (2021) all lacked structures of accountability for AI's societal impacts, making the analysis of societal impacts of AI likely to be neglected without consequences.

Also relevant in the discussion of the AI field's "architecture" is Conway's law, which states that any organization designing a system "will produce a design whose structure is a copy of the organization's communication structure" (Conway, 1968). Conway's law has been shown to be supported empirically, including in the software industry. Documented natural experiments in the software industry highlight that an organization's governance model, approach to problem-solving, and communication patterns "constrain the space in which it searches for new solutions" (MacCormack et al., 2012).

Because organizations in the industry and academia are structured quite differently, we can expect the centrality of the industry to AI R&D to leave an imprint on the structural design features of the systems created in the course of AI progress. For example, universities typically include sets of faculty with deep expertise in a well-rounded set of fields. The frequency and depth of collaborations between the faculty from different departments vary by university and individual faculty members, but at least in theory, they can always solicit an opinion of a colleague from a very different discipline working within the walls of the same

university. This is not so within industry. Not only do companies tend to be much more homogenous in terms of represented disciplines but soliciting an opinion of an academic from a different discipline often requires going through a multi-step process which includes signing a non-disclosure agreement, which many academics can be wary about, making a case for the collaboration, securing a budget, etc. These difficulties in communication flows can end up leading to AI systems created without sufficient multi-disciplinary consideration of the impact of the design choices on society and the prosperity of its members.

Conclusion

Dozens of organizations have published "Responsible AI" principles in the last few years. Those commonly include declarations of intent to make AI transparent, accountable, and beneficial to all (Fjeld, 2020). And while many organizations have begun dedicating some staff time and resources to substantiate their promises to make AI fair, explainable, and safe, very little is being done to move beyond on-paper principles when it comes to ensuring that AI does not exacerbate inequality and lead to concentration of economic power and productive capacities. On the contrary, the expectation that AI advancement will generate large left behind groups who will need to be supported by expanded retraining programs and social safety benefits like Universal Basic Income is increasingly broadly shared, adding to the troubling normalization of thinking of the current direction of AI advancement as the only one possible.

Viewing the path of AI progress as unalterable might serve certain economic and political interests or simply be a result of misguided beliefs. However, this chapter underscores that adopting this view would deny the objective importance of key factors shaping the direction of AI advancement, such as economic incentives and the policy environment the AI industry is faced with, as well as the ideas, norms, and structural features of the field. There is a growing volume of research on how policy decisions influence the direction of AI by altering economic incentives faced by AI practitioners, but much remains to be understood about the impact of norms and architecture of the AI field on its direction.

"The AI community has historically fetishized beating or replacing humans," Frank Chen wrote in Mani et al. (2021). Understanding the drivers of this "fetishized" focus is critical for our ability to govern the direction of AI. The redirection of AI towards human complementarity will not be achieved without developing ways to practically measure the impact of AI on labor demand and job quality. Transitioning to human complementarity also requires closing the major gaps in our understanding of how Lessig's four forces—legislation, market, as well as norms and structural features of the AI field—enable and fuel AI's present jobs-destructing focus.

Acknowledgments

I am thankful to Anton Korinek, Daron Acemoglu, Sophia Nevle Levoy, and Neil Uhl for their comments on the drafts of this chapter. Financial support by the Ford Foundation is gratefully acknowledged. katya@partnershiponai.org.

References

Acemoglu, D. (2010). When does labor scarcity encourage innovation? *Journal of Political Economy 118* (6), 1037–1078.

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives 33* (2), 3–30.

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society 13* (1), 25–35.

Acemoglu, D., & Restrepo, P. (2021). *Tasks, automation, and the rise in US wage inequality* (No. w28920). National Bureau of Economic Research.

Acemoglu, D., Manera, A., & Restrepo, P. (2020). *Does the US tax code favor automation?* (No. w27052). National Bureau of Economic Research.

AI Index. (2021). *AI Index Annual Report, 2021*. https://aiindex.stanford.edu/report/.

Altman, S. (2021). Moore's law for everything. https://moores.samaltman.com/.

Autor, D., Mindell, D., & Reynolds, E. (2020). *The work of the future: Building better jobs in an age of intelligent machines*. MIT Work of the Future.

Broockman, D. E., Ferenstein, G., & Malhotra, N. (2019). Predispositions and the political behavior of American economic elites: Evidence from technology entrepreneurs. *American Journal of Political Science 63* (1), 212–233.

Clemens, M. A., Montenegro, C. E., & Pritchett, L. (2019). The place premium: Bounding the price equivalent of migration barriers. *Review of Economics and Statistics 101* (2), 201–213.

Conway, M. E. (1968). How do committees invent. *Datamation 14* (4), 28–31.

Costello, B., & Karichkoff. A. (2019). *Truck driver shortage analysis 2015*. The American Trucking Associations.

Defense Advanced Research Projects Agency (DARPA). (2005). The Grand Challenge. https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles.

Defense Advanced Research Projects Agency (DARPA). (2014). The DARPA Grand Challenge: Ten years later. https://www.darpa.mil/news-events/2014-03-13.

Diao, X., Ellis, M., McMillan, M., & Rodrik, D. (2021). Africa's manufacturing puzzle: Evidence from Tanzanian and Ethiopian firms (No. w28344). National Bureau of Economic Research.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication.

Gerety, R. (2020). Expert Interview.

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Ivanov, S. (2020). ICML 2020. Comprehensive analysis of authors, organizations, and countries. https://medium.com/criteo-engineering/icml-2020-comprehensive-analysis-of-authors-organizations-and-countries-c4d1bb847fde.

Jasanoff, S., & Kim, S. H. (2015). *Dreamscapes of Modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press.

Kerry, C., & Karsten, J. (2017). Gauging investment in self-driving cars. Brookings Report. https://www.brookings.edu/research/gauging-investment-in-self-driving-cars/.

Klinova, K., & Korinek, A. (2021). AI and shared prosperity. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21), May 19–21, 2021, Virtual Event, USA. ACM, New York, 7 pages. https://doi.org/10.1145/3461702.3462619.

Korinek, A. (2019). *Integrating ethical values and economic value to steer progress in artificial intelligence* (No. w26130). National Bureau of Economic Research.
Korinek, A., & Stiglitz, J. E. (2019). *Artificial intelligence and its implications for income distribution and unemployment* (No. w24174). National Bureau of Economic Research.
Korinek, A., Schindler, M., & Stiglitz, J. E. (2021). Technological progress, artificial intelligence, and inclusive growth. IMF Working Paper.
Lanier, J. (2014). *Who owns the future?* Simon and Schuster.
Lessig, L. (1999). *Code: And other laws of cyberspace*. Basic Books.
Luckerson, V. (2016). How a big promotion at Google reveals the future of Search. https://time.com/4206532/amit-singhal-google-future-of-search/.
MacCormack, A., Baldwin, C., & Rusnak, J. (2012). Exploring the duality between product and organizational architectures: A test of the "mirroring" hypothesis. *Research Policy 41* (8), 1309–1324.
Mani, G., Chen, F., Cross, S., Kalil, T., Gopalakrishnan, V., Rossi, F., & Stanley, K. (2021). Artificial intelligence's grand challenges: Past, present, and future. *AI Magazine 42* (1), 61–75.
Nguyen, A. (2021). *The constant boss: Work under digital surveillance*. Data and Society.
Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
Partnership on AI. (2021). Redesigning AI for shared prosperity: An agenda. partnershiponai.org/shared-prosperity.
Pritchett, L. (2020). The future of jobs is facing one, maybe two, of the biggest price distortions ever. *Middle East Development Journal 12* (1), 131–156.
Raji, I. D., Bender, E., Paullada, A., Denton, E., & Hanna, A. (2020). AI and the everything in the whole wide world benchmark. ML Retrospectives NeurIPS Workshop proceedings.
Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction 5* (CSCW1), 1–23.
Reich, R. (2021). Why AI needs academia. http://bostonreview.net/forum/redesigning-ai/rob-reich-why-ai-needs-academia.
Reddy, R. (1988). Foundations and grand challenges of artificial intelligence: AAAI presidential address. *AI Magazine 9* (4), 9–21
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., & Li, F. (2015). *ImageNet Large Scale Visual Recognition Challenge*. IJCV.
Russell, S. (2019). *Human-compatible: Artificial intelligence and the problem of control*. Penguin.
Saadia, M. (2016). *Trekonomics: The economics of Star Trek*. Inkshares.
Schindler, M. M., Korinek, M. A., & Stiglitz, J. (2021). *Technological progress, artificial intelligence, and inclusive growth* (No. 2021/166). International Monetary Fund.
Seamans, R. (2021). Comments at the "Redesigning AI for Shared Prosperity" seminar held by the Centre for the Governance of AI. https://www.youtube.com/watch?v=EvcCx_qEkrs.
Siddarth, D., Acemoglu, D., Allen, D., Crawford, K., Evans, J., Jordan, M., Weyl, G.E. (2021). *How AI fails us*. Edmond J. Safra Center for Ethics, Harvard University.
Stiglitz, J. (2014). *Unemployment and innovation*. NBER Working Paper no. 20670.
Sweeney, K. (2019). Cambridge 'Star Trek' technology startup hooks millions to attack $206bn market. *Business Weekly*. https://www.businessweekly.co.uk/news/hi-tech/cambri

dge-%E2%80%98star-trek%E2%80%99-technology-startup-hooks-millions-attack-206bn-market.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6720–6731.

CHAPTER 37

PREPARING FOR THE (NON-EXISTENT?) FUTURE OF WORK

ANTON KORINEK AND MEGAN JUELFS

Introduction

In modern societies, work is both the main source of income for working-age people and the main pursuit upon which they spend their time. However, there are concerns that advances in artificial intelligence and related technologies may substitute for a growing fraction of workers, presenting significant challenges for the future of work. This chapter analyzes how to optimally allocate, or phase out, labor and distribute income in a world in which intelligent machines may increasingly become substitutes for human labor and drive down wages.

The chapter starts by laying out three distinct concerns about the future of work and analyzing them from an economic perspective. The first concern is that technological progress may be labor-saving and may reduce wages. Since the beginning of the Industrial Revolution, technological progress has benefited capital and labor in roughly equal proportions, giving rise to large increases in real wages. However, recent advances in automation have substituted for unskilled labor and have saved on labor in some segments of the labor market, reducing real wages for the affected workers. Future technological progress may reduce the wages of an ever-larger fraction of the population.

The second concern is that machines may become perfect substitutes for labor. Throughout our history, technological progress has made production more efficient, but human labor has always remained essential for the production of output. The central role of labor in production is one of the main reasons why progress has led to increasing wages. By contrast, if advanced AI and robotics can fully substitute for any type of human labor in the future, then labor will no longer be essential. We summarize predictions that labor may become perfectly substitutable within the current century. This may accelerate labor-saving progress and may mark the beginning of the end of the Age of Labor. On the positive side,

if the scarce factor of labor is no longer essential for production, economic growth may increase significantly.

The third concern is that labor will become economically redundant, which we define as the price of machines that substitute for human labor falling below the cost of human subsistence. In such a world, humans could no longer survive based on earning competitive market wages. Work would become obsolete and would cease to play the central role that it currently plays in our society.

Next we evaluate a number of objections to the three concerns we described. We discuss the notion, put forward by some, that human labor may inherently be superior to anything that machines could ever do. We describe how the articulated concerns relate to the lump-of-labor fallacy, to historical experience, and to the creation of new jobs when old jobs are displaced. We assess whether human demand is necessary for the functioning of the economy if labor is phased out. We identify a role for "nostalgic" jobs in which humans are hired merely for the fact of being human, even if machines could perform their work more cheaply and more effectively. Finally, we discuss how concerns about the economic redundancy of labor relate to the concepts of comparative and absolute advantage in international trade.

First, we will analyze how to optimally allocate work and income to maximize utilitarian social welfare as a function of the prevailing state of technology. For given labor productivity and non-labor income, an individual worker can find herself in one of three different regimes. In the first regime, the worker's productivity and non-labor income are too low and she perishes. In the second regime, which arises for higher levels of labor productivity but sufficiently low non-labor income, the agent works, first out of need, then out of choice. As her non-labor income rises, it is optimal for her to spend more and more time on leisure rather than work. She enters the third regime if technological progress increases her non-labor income or reduces the marginal product of her labor sufficiently so that it becomes optimal for her to no longer work.

In an economy with multiple agents who differ in their labor productivity, a utilitarian planner finds it optimal to allocate work to individuals solely based on their labor productivity, with more work allocated to individuals with greater productivity. If output in the economy becomes sufficiently high or the marginal product of labor declines sufficiently, it becomes optimal to phase out work, beginning with workers who have low labor productivity and job satisfaction, as they have comparative advantage in enjoying leisure. Once the income that is produced by autonomous machines is sufficiently high, it becomes optimal for humans to stop working altogether.

When we account for job amenities that workers value in addition to their pay (e.g., identity, meaning, or structure), then the picture becomes more nuanced: for workers who value such amenities sufficiently, it may be desirable to continue to work even if their labor productivity and their competitive market wages fall to zero. Work may also give rise to externalities and public goods, such as social connections or political stability. Moreover, it may entail internalities, i.e. workers may derive benefits (or costs) from work that they do not rationally internalize (e.g., from the structure to their daily lives that work provides). Such instances may lead to suboptimal decisions by workers and firms and may create a case for public policy interventions.

Next, we will discuss how to reform our economic institutions to allocate work and income when technological progress reduces wages and when work becomes economically redundant. At present, one of the main institutions to allocate both is the market. However, market failures imply that individuals have very limited access to long-term insurance markets that would protect them against adverse shocks, including against the risk of labor-saving technological progress and economic redundancy. Social insurance systems may fill the gap and provide individuals with some protection. Moreover, as our economy grows wealthier or as market wages decline, an important role of social insurance is to provide sufficient income to individuals so they are not forced into work that has low productivity and that therefore creates little social value, while using up their valuable leisure time.

Labor-saving progress makes it desirable to engage in greater redistribution. As long as labor income still plays a substantial role in our economy, our current mechanisms of social insurance form a good basis to build on by focusing benefits on those who need them most, but they should be reformed to ensure that they do not condition benefits on work. More generous transfers to workers also have multiplier effects: by reducing labor supply, they increase equilibrium wages and enable workers to demand better work amenities, which increases the incomes of workers and further raises their welfare. Moreover, as long as labor is an important source of income, it is desirable to actively steer technological progress in directions that are labor-using and increase competitive market wages so as to reduce the need for redistribution.

If labor becomes economically redundant, individuals will no longer be able to survive based on their labor income alone, and other sources of income—whether from sufficiently well-distributed capital ownership or benefit payments—are critical to avoiding mass misery and the political instability that may result from it. Because machines that make labor redundant would be able to produce unprecedented levels of output, it should in principle be easy to engage in more distribution.

As the vast majority of the population will require income support, the advantages of targeting benefits according to individual needs will decline and a basic income for everyone will become relatively more desirable. We argue that an unconditional Universal Basic Income (UBI) would be preferable to benefits conditioned on work. Recipients could still engage in work for its own sake if they enjoy work amenities, such as structure, purpose and meaning.

A role for public policy to actively encourage work only exists if work gives rise to positive externalities, such as social connections or political stability, or if individuals undervalue the benefits of work because of internalities. However, these externalities and internalities may be difficult for society to agree upon. In the long run, we conjecture that there would likely be alternative and more efficient ways of generating any positive externalities that traditionally derived from work.

Finally, we also observe that our system of taxation will have to adapt significantly as the role of labor in the economy declines. Taxes will have to be raised increasingly on factors other than labor, for example via Pigouvian taxes on activities that generate externalities, or via taxes on inelastic factors, such as land, that are not distorted by taxation. Moreover, any other factors that benefit from technological progress at the expense of labor would be good candidates for taxation.

Evaluating Concerns About the End of Labor

To begin, we will formalize and evaluate three concerns that capture the economic implications of a scenario of ever-more-intelligent autonomous machines that substitute for human labor and that drive down wages. The first concern is that technological progress may be labor-saving and reduce wages in absolute terms. The second concern, frequently articulated by technologists, is that machines may become perfect substitutes for labor. We discuss recent technological developments and summarize predictions regarding this concern. Next we analyze under what conditions this leads to a third concern, that labor becomes economically redundant. We also evaluate the main objections to predictions about the redundancy of labor. The three concerns are described using formal economic language in (Korinek & Juelfs, 2022).

Labor-Saving Progress

Since the onset of the Industrial Revolution, technological progress has led to broad-based economic growth that has, on average, benefited everybody in the economy. In particular, output and wages grew at approximately the same rate over that period—a phenomenon that economists call neutral technological progress.[1] Figure 37.1(a) illustrates neutral technological progress in a stylized diagram, in which output is depicted on the horizontal axis and labor compensation (wages w times employment ℓ) on the vertical axis. Neutral technological progress implies that labor compensation grows in lockstep with output so that the share of labor remains constant.

However, as Hicks (1932) already noted, technological progress need not always be neutral. It is well possible that future advances in technology substitute for human labor and benefit labor less than other factors, or that they even lead to declines in competitive market wages. Panel (b) of Figure 37.1 illustrates technological progress that is biased against labor:

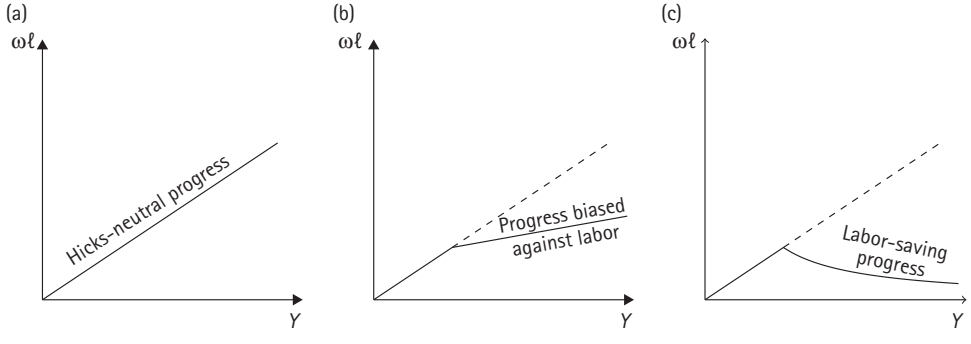

FIGURE 37.1 Varieties of technological progress

after the kink, labor compensation still grows, but at a smaller rate than output, reducing labor's share of income. Karabarbounis and Neiman (2014) document that this has been happening in the majority of countries around the world in recent decades.

Panel (c) of Figure 37.1 of the figure depicts the possibility that technological progress may be labor-saving in absolute terms, so that the compensation of labor declines even as overall economic output rises. We formalize this concern as follows:

Concern 1 (Labor-Saving Progress in Absolute Terms)

Technological progress reduces the demand for labor at given prices.

If the demand for labor declines at given prices and wages, then the equilibrium competitive wage will decline in absolute terms (as long as the supply of labor is constant or upward-sloping).

Autor (2019) documents that large categories of the United States workforce, especially lesser-educated workers, have already experienced stagnating or declining real wages in recent decades. Acemoglu and Restrepo (2022) find that the majority of these declines due to automation. What has been the fate of unskilled lower-wage workers in recent decades may turn out to be the fate of high-skilled and high-wage workers in future decades. Korinek and Stiglitz (2021a) provide several economic models that exemplify that it is well possible that technological advances may reduce competitive market wages in absolute terms even though output increases, and that this may hold even if automation is accompanied by the accumulation of additional capital. This captures what Wassily Leontief (1983) described as follows: *"the role of humans as the most important factor of production is bound to diminish—in the same way that the role of horses in agricultural production was first diminished and then eliminated by the introduction of tractors."*

Perfect Substitutability of Labor

Many predictions about the redundancy of labor are based on the premise that the human brain is at its core a computing device that processes information by transforming inputs into outputs. This premise makes it plausible that advances in hardware and software may catapult the computing capabilities of machines to the point where they may rival the human brain. When combined with sufficiently advanced sensors and actuators, machines could then perform any kind of work that humans can perform. We formalize this concern as follows:

Concern 2 (Perfect Substitutability of Labor)

Machines can substitute for any type of labor in production.

When this condition is satisfied, labor is no longer an essential input for production—given enough machines, production at any desired scale can take place without labor.

Current State of Technology

At present, artificial intelligence is clearly nowhere near possessing the ability to perform all human work. Humans cannot be substituted for and are still essential both in simple non-routine manual tasks and in higher-order cognitive tasks that require an understanding of the world, planning, and crucially, social intelligence. However, over the past decade, advances in deep learning have enabled artificial intelligence and AI-powered robots to perform a growing number of less-structured and higher-level cognitive tasks.[2]

Technological Predictions

In terms of sheer computing power, the world's most advanced computers are already roughly on par with, or superior to, the human brain. One common measure of computing power is floating point operations per second (flops), corresponding to how many arithmetic operations on real numbers a computer can perform per second. Carlsmith (2020) estimates that the computing power of the human brain can be replicated with about 10^{15} flops, given the right software. At the time of writing, Fugaku, the world's top publicly known supercomputer, was able to reach a peak performance exceeding 10^{18} flops, easily surpassing this estimate, albeit the system was reported to cost more than $1 billion. And computing capacity is expected to continue to grow for the foreseeable future.[3]

The models underlying cutting-edge AI applications are also experiencing rapid progress. Given the described hardware capabilities, the absence of a human-level general AI that we know of suggests that advances in software are somewhat lagging behind advances in hardware. However, the field is experiencing a rapid inflow of talent and funds (AI Index Report, 2021), suggesting that progress will continue unabated.

The futurist Ray Kurzweil (2005) predicts that AI will achieve human levels of intelligence in 2029 and has reaffirmed his prediction in recent years. Bostrom (2014) and Grace et al. (2018) conduct surveys of AI experts on when human-level artificial intelligence may be reached. Bostrom reports a median prediction of 2040 in a sample that included many futurists. Grace et al. report that a broad sample of AI researchers assign a 50 percent chance that humans will be technologically redundant by the early 2060s.

Economic Redundancy of Labor

Perfect substitutability does not necessarily imply economic redundancy of labor because it does not consider how costly it is to replace human work with machines. If substituting for all human labor is technologically possible, but it would cost many times more than human workers at current market wages, then it is not economically efficient to do so.

The cost at which machines can perform a given human job imposes a ceiling; that is, an upper bound on the competitive market wage of humans performing the job. As

technology advances and machines become more and more efficient, this ceiling—and, by implication, the market wages of humans—will decline. At first, jobs that are easily automated are affected, leading humans to switch to work that is more difficult to automate and therefore pays higher wages, as they have been doing for centuries. However, in the past, there were always jobs left that only humans could perform. This will no longer be the case if machines become perfect substitutes for labor. And as machines continue to become more efficient, wages may fall below the subsistence cost necessary to support a human worker so that human labor is no longer economically worth its keep. We formalize this concern as follows:

Concern 3 (Strong Economic Redundancy of Labor)

Machines are able to perform any economically valuable task cheaper than humans, valued at their subsistence cost.

Concerns 2 and 3 reflect the way in which the potential future redundancy of labor is frequently framed in technology circles. Observe that perfect substitutability of labor (Concern 2) is a necessary condition for strong economic redundancy of labor (Concern 3) but not a sufficient one. However, if Concern 2 is satisfied and a broad version of Moore's Law holds whereby the cost of human-replacing machines falls by one-half roughly every two years, then even very expensive technologies to subsitute for human labor may become affordable relatively quickly: if the cost of substituting for one unit of human labor in its most expensive use is c_m and the cost of human subsistence is c_0, then it will take $2 \log_2(c_m / c_0)$ years for Concern 3 to materialize. For example, for $c_m = \$1$ billion and $c_0 = \$1{,}000$, the process would take about 40 years.

If Concern 3 materializes, humanity would no longer add economic value—humans would require more economic resources than they are able to produce, and labor would be obsolete in the sense of becoming a dominated technology. Moreover, humans could no longer survive based on their competitive market wages alone—a stark departure from the way our societies have been organized since the onset of the Industrial Revolution—and we would have to choose between mass misery or providing a basic subsistence income, as we will explore in more detail below.

An analogous economic outcome may result even from a weaker condition. Assume that there are still some jobs that can only be done by humans (violating Concern 2) or in which humans are more cost-effective than machines (violating Concern 3). If the demand for those jobs is insufficient, then the market-clearing price of labor may still fall below the subsistence cost of humans. We formalize this as follows:

Concern 3' (Weak Economic Redundancy of Labor)

The competitive wage of human labor falls below the subsistence cost of humans.

Observe that Strong Economic Redundancy (Concern 3) implies Weak Economic Redundancy of Labor (Concern 3') in a competitive economy but not vice versa, and that Weak Economic Redundancy (Concern 3') can be satisfied even if the Perfect Substitutability

of Labor (Concern 2) is violated. (For proofs, see Korinek & Juelfs, 2022). An example that satisfies Concern 3' but does not satisfy either Concern 2 or 3 is an economy in which a small amount of labor is essential—figuratively speaking, a worker who presses the "on"-switch of the machines every morning—but in which any additional labor exhibits productivity below c_0. In such an economy, machines are not perfect substitutes, but the market clearing wage is below the subsistence level c_0.

Whereas Concerns 2 and 3 described features of the technological environment, (i.e., the supply side of the economy), Weak Economic Redundancy (Concern 3') characterizes the economy's equilibrium and therefore depends not only on technology/labor demand but also on labor supply. For example, if an economy in which labor was weakly redundant introduces a more generous welfare system and labor supply declines, it may lift the marginal product of labor so the condition no longer holds.

A critical component in both conditions for the redundancy of labor was the market price of human labor compared to workers' subsistence cost. Advances in technology may well lead to declining nominal consumer prices as production becomes more efficient. Economic redundancy would only be reached if competitive market wages decline faster than the subsistence cost. This would likely be the case if machines become even more efficient—compared to humans—at transforming factor inputs, such as energy and raw materials, into output.

Objections

Human Superiority

One common objection to the described perspective is that human labor will never become fully redundant because humans are innately superior to machines in certain domains. This belief is held firmly by many. We acknowledge that this is a possibility, but we also observe that there are no physical or economic laws that would suggest that the intelligence and dexterity of machines cannot in principle surpass their human counterparts. Human intelligence is subject to significant natural limits, for example because of natural constraints on the size of our brains. At present, these seem difficult to overcome.

Advances in the intelligence of machines have far outstripped advances in human intelligence in recent decades. If this differential progress continues, machines will eventually surpass human levels of intelligence.[4] We also note that for machines to be able to become perfect substitutes for all labor (Concern 2), they do not necessarily need to have anything corresponding to human consciousness, and they do not need to possess metaphysical attributes that many attribute to humans, such as a soul. They just need to be able to perform all tasks that have economic value in an effective manner—including tasks that involve social and emotional intelligence, which will require them to develop a sufficiently advanced theory of mind. A useful discussion of several related points is given in section 6 of Turing's (1950) famous paper on "Computing Machinery and Intelligence," in which he responds to objections to the notion of machine intelligence.

The Lump-of-Labor Fallacy and New Jobs

There have been numerous predictions about the "end of labor" in the past, most famously predictions that were based on the so-called "lump-of-labor fallacy." This is the notion that there is a fixed amount of work to be done in the economy and, by extension, a fixed number of jobs, so that automating jobs will generate persistent unemployment. The lump-of-labor fallacy is false because it fails to acknowledge the power of markets—specifically of the price mechanism—to clear demand and supply. In a well-functioning economy, surplus labor exerts downward pressure on wages, making it cheaper to create new jobs and hire workers up until the point where the market clears again. Similarly, innovation that creates additional demand for workers exerts upward pressure on wages until the market clears. From a macroeconomic perspective, what matters is not the creation or destruction of specific jobs—these are symptoms of the economy's adjustment process—but the effects of technology on overall labor demand.

Note that Concerns 3 and 3' about the economic redundancy of labor did not state that jobs would not exist or that no new jobs could be created, just that they would not pay living wages. Or more technically speaking, that the marginal product of labor would be less than the cost of human subsistence. In other words, our concern about the economic redundancy of labor is very different from the (false) concerns articlulated by the lump-of-labor fallacy.

Historic Extrapolation

Concerns about the economic redundancy of labor are in stark contrast with the historical experience since the Industrial Revolution. During the nineteenth and early twentieth centuries machines replaced our brawn so humans instead focused on more brain-intensive cognitive tasks. Since the beginning of the computer age, machines have replaced dull and repetitive structured information processing tasks, while creating new jobs for humans that involved more varied cognitive tasks and leveraged our multi-faceted human intelligence. Autor et al. (2021) document that 63.5 percent of US employment in 2018 occurred in job titles that did not even exist in 1940. Given the additional wealth generated by technological advances in each of these cases, there was also greater demand for labor, including for the newly created jobs, leading to rising wages. Aghion et al. (2019) describe in an elegant model how wages and the economy can grow continually along a balanced growth path if a constant fraction of all the tasks in which human labor is employed is automated every period—assuming that labor remains essential for the remaining tasks.

However, there are no fundamental physical or economic laws that would say that these patterns will to continue to hold going forward—it is solely based on extrapolation from the 250 years since the Industrial Revolution. Moreover, before the Industrial Revolution, Malthusian forces implied that living standards were essentially stagnant for much of humanity. This serves as a reminder that past trends will not necessarily continue in the future.

Human Demand

Another objection is that an economy could not operate without consumer demand (i.e., that humans need to earn wages so that they can afford to consume goods and services

and keep the economy going) (see, e.g., Ford, 2015). This is a fallacy—it is true that all output produced needs to be demanded by someone or something. However, (1) that demand does not have to derive from humans, and (2) even to the extent that it derives from humans, it need not be financed with labor income. On the first point, it is perfectly feasible for a thriving economy to exist in which all output is used solely for investment purposes (i.e., machines producing output to serve as input for machines) (see, e.g., Korinek, 2019). The economy would have to re-tool, for example by switching from agricultural farms to server farms that meet the demand created by machines, but there are no economic laws that would make this impossible. On the second point, human demand could also be financed by other sources of income than labor, for example by factor income such as the returns to capital, machines or land, by transfers and social benefits, or by government spending.

Nostalgic Jobs

Another objection is that even if human labor can be perfectly substituted for (Concern 2), the economic redundancy of labor will be avoided because humans will always prefer to obtain certain services from other humans rather than from machines, for reasons that we may call "nostalgic." For example, humans may prefer not to replace the services of human priests, judges, or lawmakers. In that case, strong economic redundancy (Concern 3) will not be satisfied in nostalgic jobs.

If human labor is indeed perfectly substitutable and the cost of substituting for it progressively declines, then reasoning that there will remain certain human-only nostalgic jobs relies on three assumptions: First, it requires that we can in fact tell the difference. A robot priest who has greater emotional intelligence than humans and has a more comprehensive understanding of the human psyche than a human priest may well be able to play the role of human priest quite perfectly, or intentionally slightly imperfectly so as to not give away that it is a robot. Second, it assumes that humans will still prefer services performed by other humans—even if machines have an equal or superior track record. For example, properly calibrated artificial judges may be able to make more accurate and humane judgments than humans, leaving behind the noise, discrimination and biases that have plagued our justice system (e.g., Kahneman et al., 2021). Sufficiently advanced autonomous vehicles may kill far fewer people on our roads than human drivers. It may seem brutish to insist on jobs being performed by humans in a sub-standard and inefficient way if machines can perform them better and cheaper. Third, it requires that humans earn sufficient income to spend on human services to support human jobs; that is, that the human share of income (consisting of both their labor and capital income) is sufficiently large.

If these three assumptions are satisfied for a sufficiently large number of nostalgic jobs, labor demand will remain high enough to avoid weak economic redundancy (Concern 3') and keep wages above subsistence levels. Nostalgic jobs may therefore be an important way of keeping humans employed in jobs that pay living wages. However, if the number of nostalgic jobs that survive is too small, overall labor demand may still decline to the point of triggering weak economic redundancy and pushing wages below subsistence levels.

Comparative Advantage

The concept of strong economic redundancy (Concern 3) corresponds to machines having absolute advantage in all tasks that can be performed by human labor. By contrast, the theory of international trade tells us that what matters for gainful exchange is comparative advantage, not absolute advantage—more developed nations that are technologically superior in the production of all goods can still engage in gainful trade with less developed nations because they export the goods in which they have comparative advantage (i.e., in which they are relatively more productive), and import what their trading partner has comparative advantage in.

Some invoke the principle of comparative advantage to argue that humans cannot become economically redundant when they interact with ever more intelligent machines. However, this conflates two separate questions. In trade theory, countries are assumed to possess exogenous endowments of factors such as labor that they deploy to their most efficient use, even if these factors earn a pittance. Trade theory does not usually consider whether it is actually cost-effective to maintain these factors. In our setting, by contrast, factors (including labor) are costly to maintain and producers can choose which technology to pick—under strong economic redundancy, labor is simply a dominated technology that is not worth paying for from a purely economic perspective.

Optimally Allocating Work and Income

This section lays out an economic framework to analyze how to optimally allocate work and income given the state of technology. To do so we characterize the first-best allocation of the economy. The first best describes how to allocate scarce resources in order to maximize welfare and reflects economic allocations under idealized circumstances—it assumes that we can start from an institutional blank slate (i.e., that we can design institutions or adjust existing institutions so as to achieve the described allocation of work and income). This implies that we do not restrict our thinking to conform to the institutional status quo—we just ask what is the best possible allocation that an economy could aim for given the law of scarcity, which is formally captured by a resource constraint.

We start out by considering the case of a single economic agent—an individual who has a given level of labor productivity and receives a certain amount of non-labor income (e.g., income from capital or a benefit payment). At first we abstract from any non-monetary amenities that come with work and may affect the agent's welfare, such as meaning or social connections—we describe work simply as a transaction that gives up valuable leisure time to produce output. This allows us to illustrate the fundamental forces at work in an intuitive figure. Next we examine the case of multiple individuals who differ in their labor productivity, and we analyze how work and consumption should be allocated between them in an idealized first-best world. Then we extend our framework to incorporate non-monetary amenities such as identity and meaning and analyze how this should affect the allocation of work and income. Finally, we also observe that some amenities, such as social connections, may be subject to externalities or internalities, implying that

individuals may make inefficient choices on whether and how much to work. Formal economic models and proofs for all the results described in this section are provided in the working paper version of this chapter (Korinek & Juelfs, 2022).

Work and Income for an Individual Agent

Consider an individual consumer-worker who values consumption c and experiences disutility from labor ℓ. We assume that the agent faces a subsistence level of consumption c_0 that is required for her survival. This is not a significant constraint in today's advanced economies, but it is important to include for some of the adverse scenarios that we are investigating. Assume that each unit of the agent's labor produces w units of consumption goods, where w reflects the agent's labor productivity. The units of ℓ are chosen such that $\ell = 1$ corresponds to the maximum amount of work possible; therefore w reflects the amount of output produced if the agent engages in maximum work. In a competitive market economy, the labor productivity would correspond to the agent's wage, but our analysis here is focused on the first-best and is more general than a specific market structure.

We assume that the agent also obtains non-labor income in the amount of T, which can be interpreted as income from other factors such as autonomous machines, land or capital, or as a transfer or benefit payment, and which we take as exogenous for the analysis of the single-agent case. Together, this implies a resource constraint that reflects that the agent's consumption needs to be covered by the sum of her labor and non-labor income, $c = w\ell + T$. The first-best in this economy maximizes the agent's utility subject to this resource constraint.

The agent's optimum lies in one of three regions, depending on her labor productivity w and the non-labor income T, as illustrated in the left panel of Figure 37.2. These three regions cover the whole range of possible economic scenarios that an individual agent may face as a result of the increasing automation of labor.

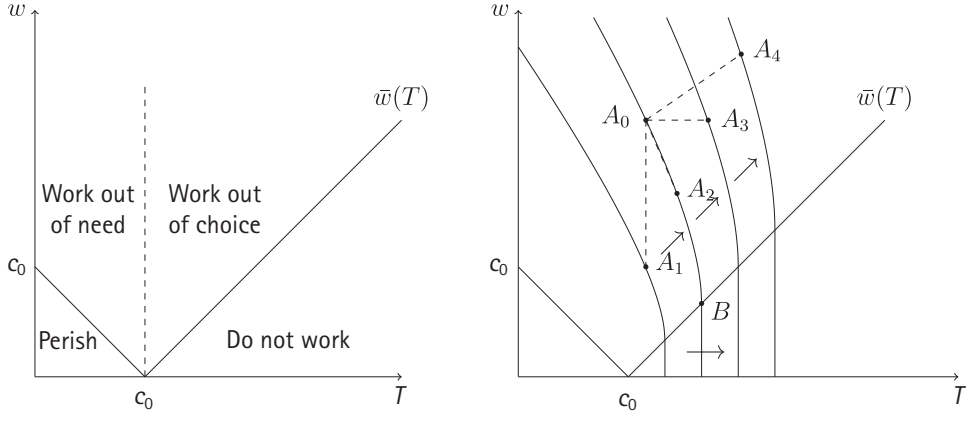

FIGURE 37.2 Regions for labor provision and iso-welfare curves

(Perish) The triangle at the bottom left of the figure is region 1, which reflects the most dystopian scenario: it occurs when the sum of the individual's potential labor income and non-labor income are insufficient to meet the subsistence level of consumption. In this region, the individual will perish. This may materialize if the individual's labor productivity diminishes as a result of ever more automation, and if the she does not have sufficient alternative non-labor sources of income to survive. The region is delimited by the minus-45 degree line representing the constraint $T + w = c_0$, which we may call the Malthusian frontier.

(Work) To the North-East of the Malthusian frontier is region 2, in which it is optimal for the individual to work. To the left of the dashed vertical line that captures $T = c_0$, it is necessary for the individual to work to cover her subsistence consumption. To the right of the vertical line, she could survive based on the non-labor income T, but it is optimal for her to work as long as her labor productivity is sufficiently high—above the threshold $\bar{w}(T)$, which corresponds to the reservation wage in a competitive economy. $\bar{w}(T)$ is an upward-sloping curve that reflects that a higher non-labor income raises the threshold that makes it worthwhile for an individual to work.

(Don't Work) The area to the right of the reservation wage curve is region 3. It captures that the individual's labor productivity is low relative to her non-labor income, making it optimal for her to enjoy her time on leisure instead of working. It would be socially wasteful for the individual to work in this region because the marginal benefit of extra consumption is declining in the level of consumption, and the extra income generated by work would not compensate the individual for her loss of valuable leisure time.

In the right panel, Figure 37.2 shows the iso-welfare lines of the agent, which are defined for regions 2 and 3. A given iso-welfare curve depicts all combinations of labor productivity w and non-labor income T for which the individual is indifferent. For example, the individual is indifferent if she moves upward or downward along the indifference curve that goes through point A_0, as any change in her labor productivity is compensated by an offsetting change in the non-labor income. This is illustrated, for example, by a movement from A_0 to point A_2. The arrows in the figure illustrate the direction in which welfare is increasing.

The indifference curves in the figure allow us to illustrate how technological changes affect welfare. In region 2, there are two distinct channels through which technological change affects individuals' welfare: changes in labor productivity, captured by vertical movements, and changes in non-labor income, captured by horizontal movements in the figure. Starting from an initial point A_0 in the figure, we illustrate several possibilities:

- Progress that reduces the labor productivity of the agent unambiguously reduces her utility, as reflected in point A_1. As we previously discussed, this is a scenario that is frequently discussed by technologists who focus on the labor-displacing properties of new technologies.
- By contrast, technological progress that increases the non-labor income T of the agent unambiguously raises her utility, as illustrated in the movement to point A_3. An increase in T arises most directly if the individual holds factors such as capital or land that become more productive as a result of the technological progress. Alternatively, it could also result from institutional changes that provide transfers to the the agent.
- Predictions of future technological progress frequently involve a combination of the two points discussed so far—a diminished role for labor but greater non-labor output. The first effect reduces utility whereas the second effect increases utility, and the overall

impact can be either. In our figure, we have illustrated the knife-edge case in which lower labor productivity is precisely offset by higher non-labor income in point A_2. Whenever the increase in non-labor income is sufficient to offset the decline in labor productivity, technological progress leaves the individual better off, i.e. on a higher indifference curve in the figure.

- The best-case scenario is illustrated in point A_4, in which both the agent's labor productivity and her non-labor income go up, increasing the agent's welfare on both counts.

Work and Income Distribution

Let us now extend our analysis to an economy in which there are many different types of agents to analyze how to optimally allocate work and income among them. We will use this framework to analyze which agents should work and how much income the different agents should receive depending on their labor productivity. This goes to the heart of the question of how to optimally distribute income and work. We continue to focus on the first-best allocation that describes the best possible allocation that an economy could aim for, given the law of scarcity.

We assume a utilitarian planner who decides how much labor and income to allocate to a set of agents who differ in their labor productivity, denoted by w^i. The planner maximizes utilitarian welfare subject to a resource constraint that captures that the agents' consumption needs to be produced using the available labor and machines in the economy.

Figure 37.3 illustrates the optimal labor allocation ℓ^i as a function of labor productivity w^i. The bold bar on the horizontal axis indicates the range of productivity levels that the agents of the economy possess, going from w^{min} to w^{max} in the example we depicted. The curves in the figure depict the optimal labor schedule of individuals as a function of their productivity. If the economy becomes wealthier and/or if the role of labor in production as reflected in the marginal product of labor declines, the optimal labor schedule shifts to the right, implying that it is optimal for individuals with given labor productivity to work less.

The figure illustrates three scenarios:

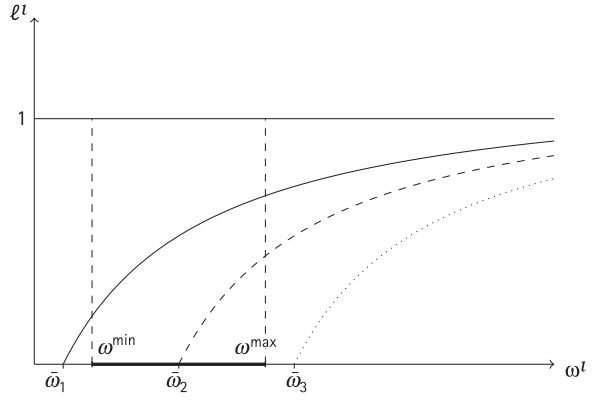

FIGURE 37.3 Optimal allocation of labor as a function of labor productivity

1. In the baseline scenario, indicated by the solid line, the reservation threshold for work $\bar{w}_1$ is below the labor productivity of the least productive agent $w^{\min}$. This implies that it is optimal for everybody in the economy to work. As the upward slope of the line illustrates, agents with greater productivity w^i are assigned a higher optimum amount of labor because their work produces more socially valuable resources. The curve is concave because the individuals' disutility of labor is rising as they work more and more.
2. The second scenario, indicated by the dashed line, corresponds to an economy that has become wealthier and/or in which labor plays a less important role in production as reflected in a lower marginal product. The reservation threshold for work $\bar{w}_2$ is now between $w^{\min}$ and $w^{\max}$, implying that work is phased out for agents with relatively low productivity, $w^i \leq \bar{w}_2$. Intuitively, it is optimal for those agents to enjoy their leisure—they have a comparative advantage in enjoying leisure and can best contribute to overall welfare by enjoying all of their time off work. For agents with productivity above the threshold $\bar{w}_2$, it is still optimal to work an amount of time that is increasing in their labor productivity; however, they work less than in the baesline scenario (solid line).
3. In the third scenario, indicated by the dotted line, the optimal labor schedule has shifted even further to the right. The scenario corresponds to an economy in which the autonomous output Y_0(i.e., the output when nobody is working) is so high that it is no longer worthwhile to employ human labor, reflected by a threshold $\bar{w}_3$ that is above the productivity level of the most productive worker in the economy. Even for that worker, it is best to spend all of her time on leisure – given the abundance of material resources, this contributes more to social welfare than her comparatively miserly labor productivity. The economy has reached full unemployment.

Regarding income distribution, a utilitarian planner places equal weights on individuals and thus allocates income equally among all agents in the first best. By implication there is no income and consumption inequality. This highlights that market allocations which tie people's income to the marginal product of their labor are sub-optimal compared to our first-best benchmark—although this is the outcome of competitive market forces, it is highly inefficient from the perspective of maximizing utilitarian social welfare. Moreover, even though such a planner would distribute income equally, there is inequality in utility—a utilitarian planner assigns more work and thus less leisure to individuals with higher labor productivity. The intuition is that the planner recognizes that individuals with low labor productivity still have the same valuation of leisure as everyone else, and it would be inefficient to make them give up their valuable leisure for work that does not produce sufficient social value.[5]

Justifying the levels of income inequality that we observe in practice requires that a planner places starkly different weights on the welfare of different individuals. To illustrate this, let us employ the commonly used CES utility function with coefficient of relative risk aversion of $\sigma = 2$ to estimate the relative marginal valuation of consumption of individuals at different consumption levels and, to be as conservative as possible, disregard the subsistence income by setting $c_0 = 0$. Meyer and Sullivan (2017) estimate that 90/10 consumption inequality is conservatively approximately by a ratio of 4:1 in the United States in recent decades (i.e., consumption at the 90th percentile was four times that at the tenth percentile).

This would correspond to the first-best choices of a planner who values individuals at the 90th percentile 16 times more than those at the 10th percentile. Comparisons across countries reveal even starker differences. For example, in several Central African countries, more than one-half of the population live below the World Bank's threshold for extreme poverty of $1.90/day, less than 1/71 of the average consumption of American households, corresponding to a welfare weight that is about 5,000 times greater.

Work Amenities

Work is not only an exchange of time against income, but it provides a bundle of non-monetary amenities that directly affect workers' welfare. Traditional job amenities include positive factors, such as benefits, flexibility, and learning opportunities, as well as negative factors, like risk exposure, pollution, and arduous working conditions (see, e.g., Rosen, 1986). As workers become more prosperous, they also increasingly care about factors such as social connections, identity, meaning, power, and self-actualization (see, e.g. Danaher, 2017). This can be interpreted through the lens of Maslow's (1943) hierarchy of needs as going from deficiency needs to growth needs.

Much of the economic discussion of the future of work focuses on how to provide income to people when they can no longer earn a living wage. However, sociologists, anthropologists, and philosophers have long observed that work also affects workers' welfare through these non-monetary amenities. This becomes especially salient when working-age people lose their jobs (see, e.g., Brand, 2015). A comprehensive discussion of economic policy for a work-less future needs to take into account the role of these factors.

We expand our analysis to account for the role of job amenities that affect an individual's welfare through channels other than consumption utility or the disutility of providing labor effort. We assume that the amenities that an individual receives from work can be captured by a scalar variable a^i, which depends on the amount worked by the individual, $a^i = \alpha^i \ell^i$, where the coefficient $\alpha^i \lesseqgtr 0$ reflects that work may give rise to either amenities or disamenities.

Under certain conditions, the benefit that an individual obtains from work—including the earnings, the disutility of labor, and the benefits or costs of the individual amenities obtained—can alternatively be derived from a compensating differential of z^i units of consumption goods. This observation is based on Rosen's (1986) principle of equalizing differences, although we look at the *total effects* of work here rather than solely the benefits derived from job amenities. This compensating differential for work is always positive, $z^i > 0$, as not working is in the choice set of the individual. By revealed preference, if the individual is working, it is because this makes her better off and allows her to earn a utility surplus compared to not working. $z^i > 0$ compensates the individual for this surplus.

However, the compensating differential may be more or less than a worker's labor income, $z^i \lesseqgtr w^i \ell^i$, depending on the value of the amenities. In traditional descriptions of the labor market that disregard amenities ($a^i = 0$), the compensating differential is always strictly less than the labor income, $z^i < w^i \ell^i$, because work requires giving up valuable leisure time. In the right panel of Figure 37.2, for example, the compensating differential for

giving up work at, say, point A_0 can be obtained by following the indifference curve to the point where it intersects with the reservation wage curve $\overline{w}$ (i.e., to point B), at which the individual ceases to work. The horizontal difference between A_0 and B reflects the transfer necessary to make the individual indifferent between working at A_0 and not working at B. If work involves disamenities to the individual, $a^i < 0$, then the individual needs to receive even less compensation to give up work, and $z^i < w^i \ell^i$ holds a fortiori.

Whenever $z^i < w^i \ell^i$, an individual whose job is displaced can be fully compensated with a benefit that is less than the wage received. Gallup (2022) reports that nearly 85 percent of employees worldwide and 65 percent of employees in the United States are not engaged or actively disengaged at work, suggesting that they fall into this category. If the job could be automated at sufficiently low cost and the surplus could be shared with the individual, overall welfare would increase.

On the other hand, for workers who receive sufficiently positive work amenities, $z^i > w^i \ell^i$ may be the case (i.e., they would need to receive a compensation that is greater than their wage earnings to give up work) because the positive amenities associated with their work more than make up for the loss of valuable leisure time. Under some conditions, it may also be possible that a worker values the amenities from work so highly that such a compensating differential does not exist. Whenever the individual derives so much utility value from work amenities that she cannot be compensated, it is optimal for her to continue to work, no matter how rich society becomes and no matter how low her labor productivity.

Let us now return to the question of how a planner would allocate work to individuals that differ in both their labor productivity w^i and the amenity value α^i that they derive from each unit of work. This introduces an additional trade off into our earlier analysis. Besides considering how much an individual's labor productivity contributes to the production of output, the planner also needs to evaluate how much an agent enjoys (or dislikes) her work.

Figure 37.4 illustrates when it is optimal for an agent to work in the space of the amenity value α^i and labor productivity w^i. The downward-sloping curve represents the agent's reservation productivity as a function of her amenity value of work α^i. If the agent's labor productivity and amenities are low, she is in the region below the curve, and it is optimal for her not to work. Conversely, if either her labor productivity or amenity values are sufficiently high, then it is optimal for the agent to work. When the amenity value is above a threshold α_0^i, then an individual's productivity no longer matters, and it is desirable for her to work even if $w^i = 0$ (i.e., even if the worker is useless in the production of output).

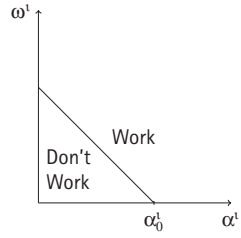

FIGURE 37.4 Optimal labor allocation with amenities

Externalities and Internalities from Work

Individual decisions on how much to work may not always be in the best interest of society, or even of the individual making the decision because work may generate rich externalities, or public amenities and disamenities, as well as internalities. Such externalities and internalities are likely to grow in relative importance if the role of work in generating income diminishes in the future.

Externalities imply that the work of one individual affects others in society not only through how much marketable output it produces but also through additional effects that are not mediated by the market. For example, the social connections engendered by work entail positive externalities as they arise not from the work of one individual in isolation but from the interactions of multiple workers. An economy with satisfactory jobs also creates positive externalities in the form of social and political stability (see, e.g., Boix, 2024). Conversely, for examples of negative externalities, the cost of commuting becomes greater the more workers are commuting because of congestion externalities. Likewise, working hard to gain greater status imposes negative positional externalities on those whose status diminishes (see, e.g., Frank, 1991).

Internalities of work occur when individuals deviate from perfect rationality in their decision-making about labor choices because they do not internalize the full effect of their choices on their own welfare. Such internalities can also be either positive or negative. For example, a hard-working individual who tells himself that work is a critical component of his happiness even though he would in fact be happier spending more time with family is experiencing negative internalities from work. Conversely, an example of positive internalities from work would be an individual who lacks structure and meaning and chooses to let his days pass by in loneliness but would in fact be happier accepting an offer for a job that provides a regular schedule and social interactions.

A utilitarian planner who decides how much work to allocate to agents needs to consider not only the agent's labor productivity, but also the true amenities derived from work (disregarding any internalities) and the externalities generated by the work. Intuitively, the planner only assigns work to an agent if their labor productivity, their true job amenities, and their contribution to aggregate amenities are sufficiently large, and the optimum amount of work for the individual is increasing in all three parameters.

Both the externalities from aggregate amenities and the internalities from mis-perceived individual amenities make the choices of individuals inefficient. To the extent that they are positive, workers drop out of the labor market too early and supply too little labor, calling for public policy interventions such as job subsidies to incentivize work. The opposite holds for negative externalities and internalities. It is optimal for the planner to subsidize labor to reflect the marginal externality as well as the marginal internality effects of an agent's labor.

One additional aspect of externalities and internalities from work that is noteworthy is that they are not properly reflected in market prices, implying that market forces may give rise to inefficient decisions by both workers and firms in the absence of regulation. For example, firms may find it profitable to automate certain jobs that generate important public goods or positive internalities because they do not internalize the losses that

society will experience as a result. Similarly, workers may happily accept a payout and stop working, even though they may end up suffering welfare losses because of the externalities or internalities from work.

Economic Institutions to Allocate Work and Income

The previous section analyzed how to allocate work and income in a "Panglossian" first-best world. This was instructive because it approaches the question from an idealized perspective that does not limit our thinking to what is possible under a specific set of institutions. However, it also risks disregarding the reasons why our institutions are set up the way they are. For example, it ignores the important role of incentives and the resulting limitations to raising government revenue and distributing income; it also ignores information problems and the associated reasons why certain markets do not exist and cannot be created in practice.

This section assesses the complementary roles played by the two main types of institutions to allocate work and income in our economy and society—by markets and by social insurance systems—when the concerns we spelled out earlier materialize.

Markets

Markets are one of the main institutions that shape the distribution of income in today's world. Economists typically start their analysis of market allocations by considering the theoretical benchmark of a free market economy in which prices are determined solely by the forces of demand and supply in competitive markets and no government intervention takes place. Just like our first-best analysis before, this is an instructive benchmark, but it is not directly applicable to the real world—as already Karl Popper (1945) observed, "the idea of a free market is paradoxical" because a functioning market requires a rich set of governmental institutions to support it, starting with institutions that protect our safety from internal and external threats to functioning judiciary, antitrust, and consumer protection institutions. It also disregards all the market imperfections that exist in the real world, ranging from externalities to other forms of missing and restricted markets.[6]

Complete markets economy

Putting aside these difficulties, we start our analysis with the benchmark case of a free market economy in which markets are complete (i.e., there are markets to trade all goods and insure against all risks, including the risks inherent in technological progress). Comparing this benchmark both with the real world and with the idealized first-best allows

us to pinpoint the shortcomings of markets and by extension the potential role for public policy to complement markets.

An economy with complete markets would—surprisingly—replicate the first-best allocation of an egalitarian planner that we previously described. Given complete markets, such an economy would also have complete risk markets, including risk markets that insure individuals and their descendents against the redistributions generated by technological progress.[7] Obviously, such insurance markets do not exist in practice. The reasons why are the same as why it is impossible to implement the first-best allocation by government fiat—perfect insurance is unattainable largely because of limits to information and incentive problems.[8] See Korinek & Stiglitz (2019) for a fuller discussion of this point.

Missing markets for long-term risk-sharing

Because the real world departs from the complete markets benchmark and markets for long-term risk-sharing are largely non-existent, individuals in a free market economy must live from their labor income and prior wealth, even when they experience stark declines in labor income. Their situation corresponds to Figure 38.2, with non-labor income T determined by the asset holdings of each individual, which derive either from inheritances or from the returns that individuals have earned on their labor and asset holdings in the past. The absence of markets for long-term risk-sharing implies that there are no incentive problems, but it assigns a large role to forces outside of the individual's control and random chance, starting from the wealth of the family an individual is born into to the luck of the draw of how much the labor market values the skill set of an individual and what returns are realized on the individual's risky investments.

Compared to the first-best allocation of an egalitarian planner, the described free market allocation without long-term insurance markets thus features large inefficiencies in the allocation of income and work: it implies that some unlucky individuals have low net worth and are forced to work even if their labor productivity is low and their work generates little economic value. Conversely, others have high net worth and no need to work even if they have high labor productivity.

If labor becomes economically redundant (Concerns 3 and 3'), then those individuals in a free market economy with insufficient asset holdings would no longer have the means to cover their subsistence cost—even if they worked as hard as they can. This would place them in the bottom left region of Figure 37.2 and leave them to perish. In the United States the net worth of the median family covers only about two years of income (Bhutta, 2020), and even less for traditionally disadvantaged groups. This implies that the majority would not survive for long in a free market economy when labor becomes redundant, leading to widespread misery and starvation. The picture would be even more dire in developing countries where wages already are closer to subsistence levels than in advanced countries and the savings of the average household are far smaller (Korinek and Stiglitz, 2021a). The described allocation would be efficient in purely economic terms—it would reflect the first-best allocation of a planner who places zero weight on individuals who succumb to starvation—it would just be incredibly cruel.

Social Insurance

Because individuals have no other source of insurance against the long-term risks of technological progress, it is natural that society has developed a social insurance mechanism that provide protection against such shocks.[9] The benefit of insurance is greater the larger the shocks that individuals are exposed to, and the consequences of labor becoming economically redundant (Concerns 3 or 3') would be among the largest conceivable economic shocks for workers.

As social insurance partially substitutes for the missing markets for long-term risk-sharing, a by-product is that it also generates an allocation in labor markets that is more efficient from a utilitarian perspective: it implies that individuals with low net worth are no longer forced into work just to survive. The more equal the distribution of income, the closer the labor supply decisions of individual workers are to the first-best benchmark of a utilitarian planner that we described, in which an individual's labor supply depends only on her labor productivity.

In the following, we first discuss the implications of labor-saving progress for social insurance. Then we examine how institutions that provide a basic income can prepare society for the economic redundancy of labor. We evaluate the pros and cons of conditioning social benefits on work and argue that unconditional benefits are likely preferable in such a scenario. Finally, we also evaluate how to adapt our system of taxation to a world of redundant labor in order to raise the revenue necessary for public spending.

Social Insurance Under Labor-Saving Progress

Much traditional economic work on optimal redistribution policy has focused on redistributing labor income from individuals with high incomes to those with low or no incomes, as labor income taxation has traditionally been the largest source of revenue for most governments (see, e.g., Piketty & Saez, 2013, for a summary). This literature centers on the classical trade off between equity and efficiency, weighing the benefit of providing to the needy against the efficiency costs of raising revenue and of distorting incentives.

A central insight of this approach is that greater inequality in market outcomes calls for more redistribution. If labor-saving progress reduces the wages earned by certain types of workers, an optimal social insurance system would increase the transfers to them. If technological progress increases overall output, an optimal social insurance system would ensure that all members of society benefit from the progress.

The social insurance systems that exist in most advanced countries are designed to target benefits to those who have the greatest needs, and under present circumstances, they do so more efficiently than a basic income for everyone would (see, e.g., Hoynes & Rothstein, 2019).[10] However, to prepare for a future in which the role of labor may decline, they should be reformed to ensure that they do not condition benefits on work, as we will

argue in more detail. In fact, transfers that are not conditioned on work also have a desirable multiplier effect.

Multiplier Effect of Non-Labor Income

As long as individuals are still working, providing transfers to workers has a multiplier effect on their income by raising equilibrium wages. Greater transfers or other forms of non-labor income induce agents to work less and, once their non-labor income is sufficiently high, to stop working altogether. In a competitive equilibrium, lower labor supply raises the equilibrium wage. This result is shown formally in the working paper version of this chapter (Korinek & Juelfs, 2022).

The multiplier effect works in both directions—if social benefits or other forms of non-labor income decline, individuals are forced to work more, which reduces equilibrium wages and lowers their incomes further. Because greater supply of labor and lower wages benefit the owners of capital, this may explain part of the political appeal to capitalists of cutting social benefits (in addition to potentially lowering their tax burden).

Role of Work Amenities

Greater income from social insurance also has a multiplier effect on the provision of work amenities. As the income obtained by workers rises—or, technically speaking, as the marginal utility of consumption of market goods goes down—the non-monetary aspects of their work become comparatively more important (see e.g., Korinek & Stiglitz, 2021b). Individuals with higher income care comparatively more about work amenities. By implication, the multiplier effect of social insurance applies to market-clearing wages as well as market-clearing amenity values.

The implications for a world with labor-saving progress are twofold: on the one hand, if individuals are not compensated for income losses that result from technological progress, they will also be forced to accept lower amenities in order to make ends meet, exacerbating the adverse effects on their utility; on the other hand, if the increase in output is shared with workers, they will demand and, in equilibrium, obtain more positive work amenities, multiplying the positive effects of social insurance on social welfare.

Steering Technological Progress

As long as labor income plays an important role in the distribution of income, Korinek & Stiglitz (2021b) make the case that it is also desirable to complement traditional forms of redistribution with a policy of intentionally steering technological progress towards labor-using rather than labor-saving innovations in order to increase market wages. Both traditional redistribution and the steering of technological progress achieve first-order gains in income distribution at the expense of second-order efficiency costs, making it desirable to employ both tools in the spirit of the theory of the second-best.

Klinova and Korinek (2021) and Korinek (2022) develop an economic framework to assess how specific technological innovations affect labor markets and inequality to assess how desirable they are from a distributive perspective. They lay out five different channels that together track the full general equilibrium effects of an innovation on labor markets. This framework enables innovators, companies, and civil society to analyze the distributive effects of their research and technology choices. It also enables public policy to consider how the policy environment for innovation affects the distribution of benefits and to actively steer innovation in a direction that benefits labor.

Economic Redundancy of Labor and Basic Income

If labor becomes economically redundant and is replaced by autonomous machines that are cheaper than human subsistence, it will affect the distribution of income in two opposing ways. First, automating labor carries the potential for vastly greater economic output. Economic growth may increase by orders of magnitude (Aghion et al., 2019; Trammell & Korinek, 2020). In principle, this should make it easy to distribute some of the surplus generated by autonomous machines to former workers to compensate them for their losses. Second, however, the ownership of capital and capital-like factors is distributed much more unequally than labor. If income from labor becomes almost negligible and no counteracting measures are taken, inequality will increase sharply. In the absence of redistribution, humans without significant financial net worth could no longer survive based on their labor income alone.

A basic income for everyone will become very appealing if labor is made economically redundant. As we observed before, in our present economic environment, targeting social benefits to recipients with low incomes makes sense because it allows our social insurance systems to focus scarce resources on where they are most needed. However, conditioning transfers on individuals' incomes will be significantly less desirable if labor becomes redundant, both because the vast majority of the population will require income support and because capital income is far more difficult to ascertain than labor income. Moreover, providing a basic income to everyone would also avoid the stigmatizing effects of income tests.

A basic income will require collective action but does not necessarily need to come in the form of transfers that are directly controlled by government. For example, Altman (2021) proposes that large corporations be required to provide equity stakes to a public fund, which would make annual distributions to all citizens. One benefit of such a scheme is that it would share the risks from technological progress widely among society, allowing everyone to benefit from economic growth and bringing us closer to the optimal insurance that would prevail under complete markets for long-term risk-sharing.

Conditioning Benefits on Work?

Many proposals for how to design a basic income when work becomes redundant suggest that benefits should be conditioned on recipients contributing to the social good, for

example by participating in public works programs or earning social credits (see, e.g., Lee, 2018; Susskind, 2020). Variants of transfers that are conditional on work include both work requirements and work subsidies for individuals, such as the Earned Income Tax Credit (EITC) in the United States. An alternative way of implementing work subsidies that relies more on the private sector is a hiring mandate for companies—for example, to require that companies need to hire a certain number of workers per million dollars in profits or otherwise pay a penalty. (If the mandate is binding, the equilibrium wage of a given job would be determined by the penalty together with the amenity value of the job, making it desirable for companies to offer enjoyable jobs.)

However, our earlier results suggest that conditioning transfers on work is only desirable if individual decisions do not deliver socially optimal outcomes because of externalities or internalities arising from work. This brings up two important questions:

1. How do we determine the most significant externalities and internalities from work? We observed that there may be both positive and negative externalities and internalities. Moreover, people differ widely in their ideas of what social benefits are delivered by work. Such disagreements make the implementation of conditional transfer schemes problematic. Letting some members of society impose their choices on others runs counter to the ideas of liberalism and freedom and may reduce welfare, counter to the intentions of the program.
2. Even if we can correctly identify them, are there more efficient ways of providing the public goods and positive externalities arising from work? Similarly, are there better ways to account for the internalities of work and provide individuals with the positive private amenities from work that they themselves do not rationally internalize, other than requiring them to work?

One way of learning about the social value of work is to look at the consequences of job loss. In a comprehensive survey, Brand (2015) reports that the negative effects of job loss are far-reaching and include several potential internalities and externalities—they range from declines in psychological and physical well-being, social withdrawal, and family disruption to reductions in the educational attainment and well-being of the children of affected workers. Crucially, however, the social context matters. Whereas unemployment fills working-age adults with dread, retirement on average leads to increases in happiness and life satisfaction (see, e.g., Fonseca et al., 2017). Well-being in retirement is also supported (1) by the narrative that workers have earned their well-deserved retirement after decades of work and (2) by a range of social customs and institutions.

This suggests two dimensions along which policy may be able to improve outcomes. First, policymakers will need to craft the right public narrative as we approach a work-less future, that work is phased out because humanity has accomplished its economic mission and can finally transition past the Age of Labor and retire into a well-deserved new epoch in which people no longer need to work. Second, it will require that we create social customs and institutions that provide well-being, identity, meaning, social connections, and any other amenities that workers currently derive from work.

This transition may take time—parents have taught their children for generations that "working hard" is an important value (IACS, 2012). Even after work has become economically redundant, it may take a new generation of citizens to fully embrace a future of human

flourishing without work, and to ensure that the positive amenities and public goods provided by work can be found in new ways. During the transition, it may be desirable for public policy to support the existence of jobs for those who want them. However, in the long run, we believe that if labor becomes economically redundant, social welfare will be maximized if humans obtain a basic income without being required to work.

A Universal Basic Income (UBI)

is the unconditional transfer program that has perhaps garnered the most public attention. A UBI is an unconditional transfer to a defined group of recipients that allows them to meet their basic needs and that is obtained no matter what economic decisions the recipients make. The unconditional nature makes it a very efficient way of distributing income.[11]

In particular, a UBI neither requires recipients to work nor actively discourages them from working—unlike social programs that are conditional on people maintaining low levels of income. If someone decides to stop working in response to a UBI, economists do not consider this a distortion or an inefficiency—in fact quite the opposite: ceasing to work would simply reflect that the recipient does not consider her work worthwhile at the given income level and market wage, and that she only performed her job out of financial duress before the UBI was introduced.

In Figure 37.2, introducing or raising a UBI would increase non-labor income T and shift the equilibrium to the right, potentially into the no-work region. If the individual obtains sufficient positive amenities from work, she can still choose to continue working, even if her market wage goes to zero.

The basic idea of a UBI is by no means new—in 1797, Thomas Paine wrote that American citizens should receive a lump sum payment upon reaching adulthood to compensate for the unequal distribution of land and wealth. In 1968, 1,200 economists signed a letter to Congress recommending a form of government assistance akin to universal basic income (Brynjolfsson & McAfee, 2014). Given the non-distortionary nature of a UBI, the proposal was supported by economists of all ideological orientations, including, for example, conservative icon Milton Friedman. President Nixon tried to implement a form of universal basic income under his Family Assistance Plan, but ultimately the plan fell apart because of concerns about its poor targeting.

In recent years, there has been a resurgence of interest in UBI as a potential solution to combat the economic effects of job loss stemming from automation and AI (Marinescu, 2017; Lee, 2018).

Preparing for the Redundancy of Labor

To future-proof our society, it would be advisable to develop institutions now to distribute a UBI that insures all members of society against the risks of labor-saving progress. We have social insurance programs that are relatively effective at targeting benefits to those who need them most, and there are few distributional benefits to a UBI. However, a modest UBI program that is designed to automatically scale up if labor-saving progress continues and the earnings opportunities of workers decline in the future would provide insurance to

workers. A good way of doing so would be to index the UBI to the growth of the non-labor part of national income. As we noted earlier, output growth is likely to go up significantly if labor becomes economically redundant. Moreover, if the labor share of income declines, the non-labor share will mechanically rise. A UBI that automatically grows in line with non-labor income would therefore constitute a social insurance mechanism for the redundancy of labor.

Making a Basic Income Truly Universal

It is worth emphasizing that for a UBI to maximize utilitarian welfare for all of humanity, the recipient pool would have to be universal in the truest sense of the word—including every citizen of the world. By contrast, most current proposals for a UBI restrict eligibility to the residents of certain geographic areas or to the citizens of certain countries.

Living up to the ambition of a UBI in the truest sense of the word "universal" would be extremely difficult to implement in practice. It would require developing significant administrative capabilities that currently do not exist, including the development of registers and record-keeping systems that minimize fraud and abuse, and payment systems that distribute the scheme. It would also require mechanisms to ensure that the citizens of autocratic states can obtain some of the benefits, for example by receiving benefits in-kind. Finally, a true UBI would also be enormously expensive, and our world does not currently have any way of raising significant amounts of revenue at the supernational level. Still, it is not unthinkable.[12]

If autonomous machines that fully substitute for human labor are developed, they will generate large economic gains. The companies or research labs developing such machines, as well as the countries in which they are developed, will obtain large windfall gains. They will be able to share some of their gains with the rest of the world, via philanthropy and via development assistance, to ensure that everyone benefits from the technological advances created by human ingenuity. It would be a good time for the world's top charities and global organizations to develop institutions that prepare our world for a future in which labor may be economically redundant.

Taxation

As the role of labor income in the economy diminishes, it will also become necessary to focus taxation on alternatives to labor (Korinek, 2020). This is in stark contrast to a trend that has taken place in most countries in recent decades and that has reduced the burden of taxation on capital income. As long as labor is the most important factor in the economy, labor taxation raises substantial revenue and is a relatively efficient way of taxation (see, e.g., Atkinson & Stiglitz, 1980). However, when labor income falls as a share of total output (as articulated in Concern 1), other sources of revenue will need to grow in importance. Moreover, if labor becomes economically redundant (as articulated in Concern 3), all taxation will have to rely on sources other than labor.

Henry George (1879) proposed that taxes on factors in inelastic supply, such as land, are excellent measures to raise revenue because an inelastic supply implies that the taxes

do not introduce distortions—the amount of land in the economy is essentially fixed, no matter what tax rate is imposed on it. Arthur Pigou (1920) observed that externalities such as pollution or congestion are ideal targets for taxation because taxes on them raise revenue while simultaneously making the economy more efficient.

Another way of identifying good candidates for taxation is to look for the winners of technological change. If technological progress reduces the returns on labor, then other factors in the economy must gain disproportionately and earn quasi-rents that were likely unexpected and can therefore be taxed without efficiency losses (Korinek & Stiglitz, 2019). If the same technological advances that devalued labor lead to large productivity gains throughout the economy, then they will generate a lot of income, making it easier to tax some of it to distribute to individuals whose role as workers was devalued.

Conclusions

Human society successfully adapted to past technological revolutions—we went from forager to farmer during the Neolithic revolution, and from farmer to worker during the Industrial Revolution. We hope that our society will also be able to adapt to the end of the Age of Labor if and when it occurs, enabling humans to enjoy their lives freed from the drudgery of having to work.

Crucially, however, the adaptation will require new economic institutions to distribute the output that is produced by autonomous machines, posing significant challenges to our existing institutions. The Neolithic Revolution led humans to establish property rights. Still, the majority of the population lived self-sufficient, largely autarkic lives. The ensuing specialization and division of labor, turbo-charged by the Industrial Revolution, implied that markets became essential for our daily needs and for the survival of the masses—an artisan could not survive without the farmer and baker who produced her daily bread. Still, this institutional arrangement allowed people to maintain an illusion of self-sufficiency, that they survived based on the market value of their own labor.

If labor becomes economically redundant, this would no longer be possible. Moreover, it would be wasteful to force humans to remain in low-productivity jobs when autonomous machines become ever more productive. We will need new economic institutions to share the output that our economy produces—institutions that do not depend on the market value of labor. At present, the share of output earned by labor is still larger than that of capital and other factors but is declining. Moreover, many experts predict that advances in AI may make it possible to build machines that are perfect substitutes for labor and that may make labor economically redundant within the current century. All this implies that it may be a good time to build such institutions now in order to future-proof our society.

Ultimately, if we establish the right economic institutions to distribute the abundant output in a world in which labor is made economically redundant by autonomous machines, we will be able to implement the dream of Arthur C. Clarke that "the goal of the future is full unemployment, so we can play."

The figures that appear in this article are licensed by Anton Korinek under the Creative Commons Attribution 4.0 International License (CC BY 4.0). To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Notes

1. In technical terms, economists call technological progress Hicks-neutral if labor and capital benefit proportionately for given factor supplies (or for constant relative factor supplies)—named in honor of Hicks (1932) who first characterized the phenomenon. They call technological progress Harrod-neutral if labor and capital benefit proportionately after the capital stock has adjusted to the new technology (or for a constant capital/output ratio). We are referring to the latter concept because we are interested in the equilibrium effects of technological progress, which typically involve an adjustment in the capital stock.
2. AI can now accomplish many economically useful tasks at levels of accuracy that either already are or soon will be super-human. Speech recognition and image recognition tasks can already be performed at human levels by state-of-the-art AI. DeepMind's AlphaFold can predict how the sequence of amino acids encoded in our genes will fold into proteins, opening new avenues for drug discovery and bioengineering. Transformers such as OpenAI's GPT-4 can produce high-quality text that is frequently indistinguishable from human-written text and even displays creativity. And machine learning has also led to rapid advances in robot dexterity, surpassing humans in many applications. At present, many of these technologies are at the lab stage, with commercial applications that share their benefits with society at large only at the beginning.
3. For more than 60 years now, the processing capabilities of cutting-edge processors have roughly doubled every two years, as predicted by a generalized version of Moore (1965)'s Law. Although there is some disagreement about the exact pace of future progress, few question that the growth in computing capabilities will continue to be exponential for some time to come.
4. Transhumanists observe that it may be possible to use human enhancement technology to enable humans to merge with machines and/or to keep up with rapid advances in machine intelligence (see e.g., Bostrom, 2014). However, such technology may be very costly and not widely accessible. Korinek & Stiglitz (2019) argue in section 6 of their paper that the long-term implications of human enhancment for inequality and for the fate of the median worker may not be very different from a scenario in which machine intelligence evolves separately from humans.
5. As a result, in a stark reversal to the outcome of our current economic institutions, low productivity agents would actually enjoy greater overall utility than high productivity agents in the first best of a utilitarian planner.
6. In particular, AI and related information goods—which are playing a crucial role in reshaping the (potentially non-existent) future of work—are non-rivalrous and therefore fundamentally inconsistent with competitive markets. Perfect competition implies that goods are traded at the marginal cost of producing an extra unit, which is zero for information goods. But this makes it impossible for producers to recoup the fixed costs that they have incurred in creating information goods (Arrow, 1962). To address this problem, our society has created intellectual property rights such as copyrights and patents (see, e.g., Shapiro, 2008). Intellectual property rights confer a state-sanctioned monopoly on the creators of information goods, which introduces one inefficiency (monopoly distortions) to mitigate another inefficiency (free-riding on information goods) and to provide incentives for the creation of information goods. However, it is important to remember that intellectual property rights are the result of government intervention and thus represent a deviation from free and competitive markets.

7. The argument requires that such risk markets existed since the beginning of time. Figuratively speaking, they would have allowed Adam and Eve to insure all their descendents against the potentially adverse effects of technological progress, creating perfect risk sharing.
8. In practice, the main instrument to share the benefits and risks of new technologies are equity markets. However, in the U.S. economy, for example, 47 percent of families do not have any exposure to the equity markets, and only 15 percent of families have direct stock holdings, with a median portfolio size of only about $25k (Survey of Consumer Finances, 2019, reported in Bhutta, 2020).
9. We employ a broad definition of social insurance here that includes all types of transfers and social protection programs in order to emphasize that all of these are designed to bring us closer to the first-best utilitarian benchmark with perfect risk markets that we analyzed earlier. This definition also includes many goods or services provided in-kind, for example for healthcare, safety, or education, which are by their nature targeted to the needs of the individual recipient.
10. To provide a numerical example, at the time of writing, paying every American a monthly UBI of $1,000 would use up almost the entire revenue of the U.S. federal government.
11. To be sure, there are two types of distortion that may be generated by UBI schemes. First, they require tax revenue which needs to be raised. Second, distortions may also arise from their eligibility criteria. For example, the Alaska Permanent Fund distributes some of the state's oil revenue to Alaska residents only. This may create incentives to acquire residency or citizenship to become eligible for a UBI. It may also increase opportunities for fraud.
12. At present, providing, for example, a $100 UBI to every single one of the 7.9bn citizens of the world would cost $790bn per year and would make a significant difference to the lives of the poorest. For comparison, the combined 2021 earnings of the top-10 corporations worldwide were $787bn (CompaniesMarketCap.com, 2022). The combined net worth of the world's 2,755 billionaires as of 2021 was $13.1tn and would provide an annual income of $786bn at a six percent rate of return (Forbes, 2022). All these numbers would rise significantly if labor becomes fully substitutable.

References

Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica 90*(5), 1973–2016.

Aghion, P., Jones, B., & Jones, C. (2019). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, and A. Goldfarb, (Eds.) *The economics of artificial intelligence: An agenda* (pp. 237–290). NBER and University of Chicago Press.

Altman, S. (2021). Moore's Law for everything. https://moores.samaltman.com/.

Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies 29*(3), 155–173.

Atkinson, A. B., & Stiglitz, J. E. (1980). *Lectures on public economics*. McGraw-Hill.

Autor, D. (2019). Work of the past, work of the future. *AER Papers and Proceedings 109*, 1–32.

Autor, D., Salomons, A., & Seegmiller, B. (2021). New frontiers: The origins and content of new work, 1940–2018. *Working Paper, MIT*.

Bhutta, N. E. A. (2020). Changes in U.S. family finances from 2016 to 2019: Evidence from the survey of consumer finances. *Federal Reserve Bulletin 106* (5), 1:42.

Boix, C. (2024). AI and the economic and informational foundations of democracy. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brand, J. E. (2015). The far-reaching impact of job loss and unemployment. *American Review of Sociology 41*, 359–375.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Carlsmith, J. (2020). How much computational power does it take to match the human brain? *Open Philanthropy Report*.

CompaniesMarketCap.com. (2022). Top public companies by earnings. https://companiesmarketcap.com/most-profitable-companies/.

Danaher, J. (2017). Will life be worth living in a world without work? Technological unemployment and the meaning of life. *Science and Engineering Ethics 23*(1), 41–64.

Fonseca, R., Kapteyn, A., Lee, J., & Zamarro, G. (2017). Does retirement make you happy? A simultaneous equations approach. In D. A. Wise (Ed.), *Insights in the economics of aging* (pp. 339–372). University of Chicago Press.

Forbes (2022). World's billionaires list 2021. https://www.forbes.com/billionaires/.

Ford, M. (2015). *The rise of the robots*. Oneworld Publications.

Frank, R. H. (1991). Positional externalities. In R. Zeckhauser (Ed.), *Strategy and choice: Essays in honor of Thomas C. Schelling* pages (pp. 25–47). MIT Press.

Gallup (2022). How to improve employee engagement in the workplace. Technical report.

George, H. (1879). *Progress and poverty*. Appleton.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research 62*, 729–754.

Hicks, J. (1932). *The theory of wages*. Macmillan.

Hoynes, H., & Rothstein, J. (2019). Universal basic income in the United States and advanced countries. *Annual Review of Economics 11*(1), 929–958.

IACS (2012). Culture of American families: Tables from the national survey. Technical report, Institute for Advanced Studies in Culture.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Karabarbounis, L., & Neiman, B. (2014). The global decline of the labor share. *Quarterly Journal of Economics 129*(1), 61–103.

Klinova, K., & Korinek, A. (2021). AI and shared prosperity. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21*).

Korinek, A. (2019). The rise of artificially intelligent agents. *Working Paper*, *University of Virginia*.

Korinek, A. (2020). Taxation and the vanishing labor market in the age of AI. *Ohio State Technology Law Journal 16*(1), 244–257.

Korinek, A. (2022). How new technologies affect labor markets and inequality: An impact assessment. Brookings CRM *Working Paper*, Washington, DC: *Brookings*.

Korinek, A., & Juelfs, M. (2022). Preparing for the (non-existent?) future of work. *NBER Working Paper*, 30172.

Korinek, A., & Stiglitz, J. E. (2019). Artificial intelligence and its implications for income distribution and unemployment. In A. Agrawal, J. Gans, & A. Goldfarb, (Eds.), *The economics of artificial intelligence: An agenda* (pp. 349–390). NBER and University of Chicago Press.

Korinek, A., & Stiglitz, J. E. (2021a). Artificial intelligence, globalization, and strategies for economic development. *NBER Working Paper*, 28453.

Korinek, A., & Stiglitz, J. E. (2021b). Steering technological progress. *Working Paper* Columbia University.

Kurzweil, R. (2005). *The singularity is near*. Viking.

Lee, K. -F. (2018). *AI super-powers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Leontief, W. (1983). National perspective: The definition of problem and opportunity. In *The long-term impact of technology on employment and unemployment: A National Academy of Engineering Symposium*. National Academies.

Marinescu, I. (2017). No strings attached the behavioral effects of US unconditional cash transfer programs. Technical report, Roosevelt Institute.

Maslow, A. (1943). A theory of human motivation. *Psychological Review 50*(4), 370–396.

Meyer, B. D., & Sullivan, J. X. (2017). Consumption and income inequality in the U.S. since the 1960s. *NBER Working Paper*, 23655.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine 38*(8), 82–85.

Pigou, A. C. (1920). *The economics of welfare*. McMillan.

Piketty, T., & Saez, E. (2013). Optimal labor income taxation. In A. J. Auerbach, R. Chetty, M. Feldstein, and E. Saez (Eds.), *Handbook of public economics* (vol. 5, pp. 391–474). North Holland.

Popper, K. (1945). *The open society and its enemies*. Routledge.

Rosen, S. (1986). The theory of equalizing differences. In Orley C. Ashenfelter and Richard Layard (Eds.), *Handbook of labor economics* (pp. 641–692). North Holland.

Shapiro, C. (2008). Patent reform: Aligning reward and contribution. In A. B. Jaffe, J. Lerner, & S. Stern, (Eds.), *Innovation policy and the economy* (vol. 8, pp. 111–156). University of Chicago Press.

Susskind, D. (2020). *A world without work: Technology, automation, and how we should respond*. Metropolitan Books.

Trammell, P., & Korinek, A. (2020). Economic growth under transformative AI. *GPI Working Paper*, *8–2020*.

Turing, A. (1950). Computing machinery and intelligence. *Mind LIX* (236), 433–460.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2021). *The AI Index 2021 Annual Report*. Human-Centered AI Institute, Stanford University.

SECTION VIII

DOMESTIC POLICY APPLICATIONS OF AI

JOHANNES HIMMELREICH

CHAPTER 38

ARTIFICIAL INTELLIGENCE FOR ADJUDICATION

The Social Security Administration and AI Governance

KURT GLAZE, DANIEL E. HO, GERALD K. RAY, AND CHRISTINE TSANG

Introduction

The use of artificial intelligence (AI) in adjudication is controversial.[1] Popular press accounts are chock full of alarmist accounts of "robo-judges" replacing humans (e.g., Sayer, 2016). How can the adjudicatory process, which is fundamentally concerned with tailoring law to circumstance, rely on *automated* decisions? France has gone as far as to criminalize the use of judicial analytics (Tashea, 2019). The skepticism of the use of analytics in adjudication echoes an earlier wave of critiques of quantitative approaches to judicial behavior (Edwards & Livermore, 2008; Tushnet, 1980).

Yet there is one adjudicative system in the United States that has been able to use AI to help its judges and attorneys make core adjudicative decisions: the Social Security Administration (SSA) Disability Program. In its most ambitious form, SSA has developed and deployed an automated AI system that enables adjudicators to check draft decisions for roughly 30 quality issues, addressing long-standing questions about the accuracy, consistency, and speed of case processing (Ames et al., 2020).

How did this come to be? How did "the largest adjudication agency in the western world" (*Barnhart v. Thomas*, 540 U.S. 20, 28–29 (2003)) overcome well-known structural challenges in the public sector to climb this "Agency Everest" to become a poster child for AI innovation in government? In this chapter, we tell the story of how SSA overcame significant roadblocks to develop and implement AI use cases and draw policy implications for AI innovation. This account contributes to three central questions in administrative adjudication, AI governance, and public administration. First, the story is important for understanding how to develop AI in large organizations and specifically in the challenging context of mass adjudication, such as immigration adjudication, Medicare appeals, veterans

benefits determinations, and patent examination (Ho, 2017). The SSA case study illustrates the process by which AI can be deployed to advance, not undermine, due process goals in adjudication.

Second, the SSA case study has broad lessons for AI governance in the public sector. A recent executive order establishes guidance for federal agencies regarding the adoption of AI and its use in the delivery of services.[2] It commits federal agencies to accelerate the adoption of AI in ways that will modernize government and cultivate public trust in AI. Likewise, the U.S. National AI Initiative anchors itself around principles for trustworthy AI. The SSA case holds important lessons for turning AI governance principles into practice in one of the most contested terrains. Most importantly, it shows the importance of what we call "blended expertise"—i.e., expertise at the intersection of domain and technical knowledge—to identify, develop, and test new AI-powered innovations in a way consistent with governing law and policy (Engstrom et al., 2020; Engstrom & Ho, 2020).

Third, our case study also informs our understanding of public administration and innovation (Bovens & Zouridis, 2002; Busch & Henriksen, 2018; Cinar et al., 2019). This literature has examined conditions for innovation, the technological transformation of street-level bureaucrats to "screen-level" or "system-level" bureaucrats (Bovens & Zouridis, 2002; Bullock et al., 2020), and the impact of AI systems on public administration (Criado et al., 2020; Young et al., 2019). Our case study confirms the importance of where the core organization responsible for innovation is located (Moldogaziev & Resh, 2016) and the involvement of the end users of a system in development (Criado et al., 2020). While some have theorized that AI is most impactful for tasks with low levels of discretion (Bullock et al., 2020; Young et al., 2019), we examine the use of AI to improve a highly complex, discretion-laden area of public administration: adjudication.

This chapter proceeds as follows. Section II will provide background on the SSA Disability Claims System. Section III discusses the historical challenges in the accuracy of decision making. Section IV discusses the groundwork of electronic case management, case analytics, and policy development that enabled the agency to pilot AI use cases. Section V discusses how the SSA grappled with technology governance and innovation barriers when first piloting these use cases. Section VI discusses AI applications that SSA has developed for their adjudication program. Section VII draws lessons for AI governance from SSA's experience, focusing on data infrastructure, leadership support, blended expertise that spans domain and technical knowledge, the software environment, and continuous iteration. Section VIII concludes.

The Social Security Disability Claims System

Under the Social Security Act, SSA provides disability benefits to individuals between the ages of 18 and 65 who meet the insurance requirements of the program and are unable to work because of a disability. The system for making disability determinations at SSA comprises the largest adjudicatory system in the United States (Ames et al., 2020), which paid over $200B to 18 million Americans in Fiscal Year (FY) 2020. A successful claimant

will typically receive around $270,000 in lifetime benefits, plus Medicare coverage (Gelbach & Marcus, 2016).

The Social Security Act defines "disability" as the "inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment which can be expected to result in death or . . . last for [at least] twelve months" (42 U.S.C. § 423(d)(1)(A)). SSA uses a five-step sequential evaluation process for ascertaining disability:

1. Is the claimant engaged in substantial gainful activity? If yes, the claim is denied.
2. Does the claimant have "a severe and medically determinable physical or mental impairment" or combination of impairments lasting of sufficient duration? If no, the claim is denied.
3. Does the impairment meet the severity of roughly 200 listed impairments? If yes, the claimant is disabled.
4. Does the claimant retain "residual functional capacity" to perform any past relevant work? If no, the claimant is disabled.
5. Is the claimant able to perform any other work existing in significant numbers in the national economy, based on the claimant's residual functional capacity, age, education, and work experience? If no, the claimant is disabled.

The adjudicatory process of the disability claims system works in four stages. First, the State Disability Determination Service (DDS) processes the application and makes an initial determination about whether the applicant is disabled (20 C.F.R. §§ 404.900, 902). Second, a claimant can request reconsideration of this initial determination and submit additional evidence (20 C.F.R. §§ 404.904, 913). Third, after receiving the results of the reconsideration, a claimant can request a hearing with an Administrative Law Judge (ALJ) for *de novo* review of their claim and submit additional evidence not available at the time of prior proceedings (20 C.F.R. §§ 404.915, 916, 935). Roughly 1,500 ALJs across 162 hearing offices preside in these due process hearings. Medical and vocational experts are sometimes called upon to provide opinions (HALLEX I-2-5-34, I-2-6-74), and the claimant has the right to appear at a hearing (20 C.F.R. § 404.930). Last, the Appeals Council, which together with its staff comprises the Office of Appellate Operations (OAO), considers appeals of ALJ decisions and represents the final level of appeal within the agency. Although claimants may further pursue their claims in federal district court, the vast majority of decisions are resolved internal to SSA. Most of due process, in that sense, functionally plays out inside the agency (Mashaw, 1985, 1973).

Historic Challenges

Because determinations are both factually and legally complicated, the disability claims system has faced serious challenges. A seminal study found that wide disparities between ALJs appeared driven by subjective factors (Mashaw et al., 1978), and such disparities continue to present day (Engstrom & Ho, 2020). Between 2008–2019, most claimants waited over a year for an appeal to be resolved and nearly 110,000 applicants died during that period prior to receiving a final disability decision (Government Accountability Office, 2020).

The high volume of cases has made quality improvement and the incorporation of federal court opinions very challenging (Ames et al., 2020; Gelbach & Marcus, 2017).

While procedural due process mandates "accuracy" of decisions, the accuracy and quality of decision-making is hard to verify. Employees are trained to follow agency policy, but many policies are open to interpretation. This problem can be compounded for claims involving medical evidence, which is also subject to the interpretation of medical service providers. The volume of casework and the need to adjudicate quickly can further affect quality, as individual adjudicators develop heuristic shortcuts that they employ in case processing, which may lead to incorrect outcomes in some cases.

Insufficient training, decisional (heuristic) shortcuts, the lack of a clear quality standard, and gaps in or loosely written policy guidance all contribute to the high variability in decision-making. Quality is often measured with a bottom-line approach, meaning whether the ultimate outcome is *acceptable* within the statutory and regulatory scheme (Ames et al., 2020). For example, SSA's hearings process uses the term "legally sufficient" to describe a decision of adequate quality, with the hearing level procedural manual specifying that decisions should be accurate and legally sufficient (HALLEX I-2-8-1). Other markers of quality, such as remand and grant review rates of hearing level decisions, can themselves be impacted by the variability of the reviewers of the decisions, and may do little to address the variability in hearing decisions (Gelbach & Marcus, 2017; Mashaw, 1973).

Foundational Infrastructure

In light of these challenges, how did SSA become the pioneer for pathbreaking use cases of AI in adjudication? We document here some of the foundational steps—data infrastructure, policy clarification, and analytics personnel—that facilitated the adoption of AI tools. Many of the steps were taken long before SSA considered AI applications. While these steps gave SSA a leg up when considering AI use cases, we emphasize that these are not necessary predicates for considering all types of AI. Many pilots can be conducted in parallel to IT modernization but taking these steps with downstream AI use cases in mind can be highly beneficial.

Digitizing and systematizing workflow

The efforts of the SSA to systematize and digitize its core workflows through electronic systems created highly valuable data and data infrastructure for later AI applications. Between the 1990s and 2000s, SSA implemented several electronic case management systems (eCMS) to organize its case activities and developed electronic folders to store digitized copies of claim evidence related to each claim. It also built out case analysis tools to structure and record staff notes and analysis about a claim's merits.

In the mid-2000s, early efforts to build tools to improve the quality and consistency of adjudication emerged. SSA created an electronic case analysis tool (eCAT), an electronic questionnaire that guided adjudicators at the DDS through policy compliant pathing to reach a disability determination, while capturing structured data. At the same time, SSA

built a similar tool directly into the eCMS for the Appeals Council. Known as the Appeals Council Analysis Tool (ACAT), the tool used a similar electronic form with questions to guide adjudication while capturing structured data about disability decisions. Appeals Council Members used the existing structure of SSA policy to develop a decision tree which mapped the policy compliant pathing to approximately 2,000 types of decisions possible under the sequential evaluation process. This decision tree was used in creating ACAT (Ray & Lubbers, 2014).

Prior to ACAT, analysts assisting the Appeals Council Members used a wide variety of forms for disability case analysis. The bulk of the analysis consisted of a written description of potential remandable issues and a recommended course of action for Appeals Council Members. The free-form nature of these analyses may have contributed to variability in adjudication.

The Appeals Council attempted to address this variability issue by more narrowly defining a quality decision as one that is factually accurate, procedurally adequate, supported by the record, and policy compliant. The Appeals Council used ACAT to bring a more uniform structure to this analysis and, importantly, was able to obtain access to the data captured in ACAT and the new electronic case management system. The Council began analyzing this data to provide feedback to Appeals Council Members to encourage greater consistency in adjudication. Ultimately this effort provided a clear baseline for evaluating the quality of disability analysis and later proved important for the development of the quality flags used by the Insight tool described below.

Structured policies and procedures

Over the years, SSA developed an increasingly structured process for evaluating disability claims. For example, court decisions on the use of vocational expert evidence led the agency to develop a series of Medical-Vocational guidelines that directed a conclusion of "disabled" or "not disabled" based on factors such as the claimant's residual functional capacity, age, education, and prior work experience. These guidelines, known colloquially as "the Grid Rules," take administrative notice of information contained in the Dictionary of Occupational Titles (20 C.F.R. § 404.1569, and Appendix 2 to Subpart P of Part 404), reducing dependence on vocational experts and improving consistency of adjudication. SSA also developed a sequential evaluation process for disability claims described in Section II. Class action litigation in the 1980s led to other refinements in policy structure, particularly related to the evaluation of subjective complaints and medical opinions.

To further refine SSA's policies and procedures, OAO began taking steps to leverage their access to data and reliable data infrastructure to improve the quality and consistency of adjudication. For instance, ACAT included data about why cases were remanded by the Appeals Council and federal courts. OAO generated data visualizations, such as heat maps that applied color-coding to identify trends and easily observe the frequency of error types across different hearing offices (see Table 38.1). This data-backed approach allowed OAO to quickly identify the policies generating the most substantial adjudicatory errors. Executives also addressed several circuit and district court judicial conferences and described the differences in district court behaviors to the judges, most of whom had never seen detailed data about their adjudications.

Table 38.1 Caption: Sample heat map indicating the cited reasons for remands (in rows) of cases across district courts (in columns). The numbers represent the percentage of cases remanded by each district court based on the reason cited. Darker shading indicates a more commonly cited reason for remand. The acronym "RFC" is used in place of the term "residual functional capacity," an SSA-used term describing the capabilities of a disability applicant after consideration of the limitations caused by their medical impairments. Sample heat maps to spot decisional errors

Cited Reason	WAWD	NYED	CACD	FLMD	NYSD	ILN D	ARWD
Opinion Evidence Evaluation and Residual Functional Capacity							
Consultative Examiner—Weight Accorded Opinion Not Specified	0.2	0.1	0.1	1.3	0.2	0.2	0.0
RFC—Other	1.2	0.4	1.3	1.3	0.5	1.9	0.7
Treating Source—Opinion Not identified or Discussed	2.9	1.7	2.4	3.6	4.6	2.7	1.7
Treating Source—Opinion Rejected Without Adequate Articulation	14.6	6.2	19.4	17.1	15.7	16.9	7.9
Treating Source—Weight Accorded Opinion Not Specified	0.0	0.6	0.4	3.0	0.7	0.7	0.0
RFC—Manipulative Limitations Inadequately Evaluated	1.3	0.1	0.7	2.5	1.9	1.7	4.2
RFC—Mental Limitations Inadequately Evaluated	2.5	1.1	2.4	4.1	4.6	3.9	7.2
RFC—Exertional Limitations Inadequately Evaluated	0.9	1.1	1.3	1.8	3.0	3.1	6.7
Non-Medical Source—Opinion Not Identified or Discussed	0.7	0.1	0.1	0.3	0.0	0.5	0.2
Non-Examining Source—Opinion Not Identified or Discussed	1.1	0.1	0.7	1.0	0.2	1.5	0.0
Consultative Examiner—Inadequate Support/Rationale for Weight Given Opinion	13.1	1.4	5.6	4.8	3.5	1.9	1.7

The availability of this type of granular data enabled "focused reviews" of critical problem areas and informed training programs for adjudicators. Focused reviews provided specific information about how an ALJ evaluated the evidentiary record and applied agency policies and procedural guidance and helped OAO staff identify training issues related to the misinterpretation or misapplication of policy guidance. OAO found that nearly all adjudicative errors were inadvertent. The errors were likely caused by heuristics that adjudicators adopted as shortcuts to skip aspects of policy-compliant pathing and generally still reach a policy-compliant result. Occasionally, however, the shortcuts resulted in noncompliant decisions. Training materials on reasons for remand were made available online and enabled adjudicators to self-study and close gaps in their knowledge.

OAO staff also addressed issues prone to misinterpretation where policy guidance was either unclear or insufficiently precise. OAO staff proposed policy and procedural changes that would aid adjudicators in understanding and correctly applying the policy. Staff researched the history and legislative intent of laws, background information and historical changes to regulations, as well as memoranda, legal opinions, and procedural guidance. OAO also undertook hundreds of changes to procedural guidance for hearings and appeals operations (in a manual known as the HALLEX). Clarifying its policies and procedures was important to SSA's development of rule-based AI that required such specificity for its structure and user adoption. Combined, these policy efforts laid the groundwork for some of the AI features SSA later developed.

Overcoming Organizational and Personnel Barriers in Technology Governance

The path of AI innovation seen at the SSA's Office of Appellate Operations was possible not because of SSA's existing institutional structure, but rather despite it: OAO embarked on a campaign of "stealth innovation."

As documented in public sector innovation scholarship, agency culture and climate can often impede innovation (Cinar et al., 2019). Securing support for new ideas and projects can be challenging. At the highest levels, there are many competing requests for limited resources, and most resources are allocated to maintain existing processes. Resource constraints cause new project ideas to be deferred or dropped altogether unless executives are persistent across multiple budget cycles. Even if an executive can get a project off the ground, such as by explaining how the project might improve staff productivity or the timeliness or quality of work, an immediate focus on measurable and reportable results can limit the development and exploration of new and longer-term opportunities that might flow from the initial concept.

After early requests for staffing were rebuffed, OAO decided to neither advertise the efforts it was undertaking nor seek resources for these projects. This approach provided OAO with the latitude to fully explore a range of ideas without being worried about the results, provided OAO met business objectives.

Without budget or additional headcount, OAO first set about freeing up resources for their work by improving the productivity of its existing team. OAO established numeric productivity performance standards for the professional staff, publishing internal branch goals for productivity and timeliness, and reorganizing case flow. An Appeals Council Member identified performance measures for 12 categories of work activities. The performance measures were based on the types of actions taken and hours worked over a two-year period, and baselines were set for successful and outstanding performance levels based on these measures. OAO also developed and implemented continuing learning techniques, reducing the time for trainees to become fully productive from 18 months to five months. These efforts yielded a significant rise in the productivity of the staff, from 94 case dispositions per staff member in FY 2009 to 146 in FY 2013, to 161 dispositions per staff member in FY 2017.

As productivity rose, OAO gained some latitude to branch out and move existing resources into more ambitious projects. OAO started slowly by borrowing the services of one data scientist from another SSA component, actively recruited lawyers with data science backgrounds, and developed a summer intern program to temporarily hire law students to assist with the development of training material. OAO then reprogrammed staff to address policies and procedures and expand training, and eventually added more staff to data analytics efforts.

It was only after demonstrating some success—by improving productivity and timeliness and developing compelling data visualizations—that OAO began to describe externally what it had done. OAO then sought permission for a small budgetary allocation to expand its quality assurance efforts and supplemented the allocation by also reprogramming much of its attrition hiring into quality assurance. The quality assurance staff assisted with data analytics, potential fraud investigation and analysis, and policy analysis and formulation projects. Ultimately, OAO deployed roughly six percent of its staff to long-range, exploratory analytics projects not previously undertaken by the agency. They also performed the critical, labor-intensive work required to develop modern AI tools for adjudication, like data labeling, which required high-caliber subject matter experts well-versed in the law. Repurposing trained analytical staff in this way was risky, particularly in light of increasing caseloads and staff turnover (due to a large retirement wave), but OAO was able to achieve its operational goals and continually reduce the size and age of its pending workload during this time.

OAO also sought to develop internal capacity that traversed technical and domain specific knowledge, an approach somewhat inconsistent with agency norms. Agencies tend to organize around function, with component parts of the agencies specializing in performing certain tasks, which can lead to a silo effect that can be detrimental to agency functioning and innovation in particular (Cinar et al., 2019). Performance plans, promotion paths, and bargaining unit agreements all pressure individuals to stay in their lane to succeed and advance. This dynamic of narrow focus operates not only at the component office level but also at the person-level. Agency employees are provided with a specific position description outlining the duties of their job. Work outside of this scope generally can only be performed for short periods of time, as part of an official detail to perform those duties. Often such details must be announced, and employees are selected through an open competitive process in accordance with bargaining unit agreements.

OAO took a few important steps. First, OAO focused on hiring and cultivating individuals with blended expertise. To overcome restrictions that limited OAO to hiring attorneys, OAO identified existing and new attorneys with backgrounds in statistics, mathematics, econometrics, computer science, and adult education. OAO assembled these individuals into teams to address questions of policy, training, data analytics, and the innovative use of technology. These cross-cutting teams acquired domain expertise by adjudicating cases, while at the same time developing ideas on how to deploy analytics more effectively. In addition, OAO borrowed the services of several SSA data scientists and operational research specialists to assist in cleaning, summarizing, classifying, and analyzing the data captured by analysts and adjudicators using ACAT and other data sets as they became available.

Second, personnel were given substantial space to explore a range of use cases. Such leeway ultimately enabled the group to strategically sequence use cases that would best align with the mission of SSA: namely, to improve caseload production and the quality of

decisions issued. Federal agencies like SSA often view IT projects as largely finite in development effort and cost: a large, heavily resourced development team executes a project roadmap to deliver a feature-complete product. After delivery, a much smaller team maintains the system, capable of making necessary changes but not building significant new features. Federal budgetary, contracting, and IT reporting requirements may reinforce this approach (Rubenstein, 2021). Projects with less clear end states may be viewed as running counter to these norms. And AI projects that support or make decisions are much more likely to require substantial and continuous attention and evolution as policies, business processes, and even operational norms evolve. Put differently, the AI system may require continuous resources akin to a human staff that require continuous training during their tenure (Casado & Bornstein, 2020).

Third, OAO developed some more open-ended position descriptions that provided managers with more flexibility for assigning duties, particularly among the clerical and support staff positions. The combination of improving technology and more efficient support staff enabled the redeployment of some positions into analytical jobs, which also improved productivity and performance. The flexibility in some of the position descriptions also provided the opportunity to deploy employees into policy analysis work, which included analysis of data and information. Additionally, executives and managers worked with bargaining unit representatives to creatively extend details by keeping the employees active in normal job duties part of the time and when performing overtime work.

The result of these efforts was to embed data scientists and attorneys with technical facility in an operational environment. The initiative was not without risk—as the use of staff time could be seen as a drain on case production—but OAO made a bet that data analytics would, in the long run, improve the quality, consistency, and timeliness of disability adjudication, reducing errors and the reworking of cases.

Despite OAO's success, there is still no clear promotional path to encompass the types of activities performed by blended teams like those it developed. The success of such teams may be poorly recognized outside of the highest executive channels, while line managers struggle with managing employees who may have been responsible for large business value in an area outside of the duties normally performed in that position. Promotional paths available to an employee working within specific position descriptions may lead the employee away from the very work they so successfully performed, while work more closely aligned with what they accomplished may be out of their reach or fall within another siloed component over which the employee's manager has little leverage in the promotional process. In short, inventing around barriers is not ideal.

AI Use Cases

Structured learning for workload management

With a strong foundation of data infrastructure, policy, and personnel, OAO made a number of focused bets designed to use data science to improve the efficiency and quality of adjudication. One prototype was based on the notion that it would be easier for adjudicators to consider cases involving similar issues together. If similar cases could be assigned in batches,

adjudicators might recognize the similarities in issues and require less time researching the relevant regulations and policies. The question was therefore whether a simple reordering of the assignment of work could lead to significant gains in efficiency.

To identify cases with similar characteristics without first reviewing the case files, Appeals Council Members and OAO staff worked closely with data scientists to develop a clustering analysis of the pending workload. This was accomplished by using structured data from ACAT and other hearing office level data to train algorithms that sorted cases into small batches with similar characteristics. Because the agency has interpreted the Administrative Procedure Act to preclude specialized units from processing cases by case type, the cases were worked in the usual manner by the same employees who otherwise would have worked them, just not in the same order. This project, though implemented for a limited time, appeared to reduce processing time and the need to rework erroneous cases.

OAO also worked to develop a naive Bayes supervised learning model of pending hearing level workloads. The analysis was designed to estimate the probability of an award of benefits based solely on certain characteristics found in the metadata captured in the hearing level eCMS and ACAT. The project was later extended to identify claims dismissed for procedural reasons that otherwise would have resulted in an award of benefits. Probability of outcomes were predicted but were not shared with adjudicators so as to not prejudice outcomes. Cases with higher probabilities of allowance generally were worked before other cases to speed processing for claimants most likely to be found disabled. SSA officials reported that the model overestimated the number of cases likely to be allowed (10 percent of cases as compared with the average fully favorable rate of 2.5–3 percent), but was useful in moving likely allowances ahead in the pending workload queue.

Other parts of SSA have similarly begun to integrate machine learning. SSA created a computational (Hadoop-based) resource to store and analyze data quickly. It also developed the Quick Disability Determination (QDD) process which uses a predictive model to identify cases involving one or more impairments that usually result in disability. The QDD process enabled the agency to skip resource-intensive hearings when cases were likely to result in an award.

AI to support adjudication: Insight software

Entrepreneurs within the agency with both subject matter and technical expertise were critical in spurring AI innovation by devising an ambitious suite of decision support tools known as Insight.[3] Adjudicative staff at the hearings and appeals levels of adjudication work with written decision documents. Staff generate the basic structure of the decision using a template system and then manually add substantive findings and rationale. For decades, the well-established practice was for staff to work on these decisions independently in assembly line fashion. For instance, an ALJ would prepare instructions directing the content of a decision, which would then be handed off to a decision writer to independently transform the document into a draft decision, who then handed off the completed draft back to the ALJ for independent review.

SSA leveraged Insight to improve this siloed approach. With a click, staff can now analyze their decision document and receive alerts on potential quality issues as they work, plus receive a variety of case-specific reference information and tools enriched by what

Insight found in the decision's content. At the hearing level, staff use Insight to analyze draft decisions, enabling them to evaluate and react to Insight's quality feedback prior to issuance. At the appeals level, staff use Insight to analyze issued hearing decisions under their review, helping to ensure they identify and evaluate all potential quality issues prior to making a recommendation to appellate judges. Importantly, Insight is explicitly designed only as an assistive tool—it does not decide any element of a decision nor advise any specific remedy to potential quality issues. Rather, staff are trained and even explicitly reminded in the interface that Insight's content is to serve only as a jumping off point for further analysis.

Insight's features require several AI technologies to function. First, Insight applies natural language processing (NLP) to extract information from the written decision, such as details of its findings and rationale. Insight then retrieves existing structured data about the case and claimant from workload systems (e.g., claimant claim history and biographical data, etc.). Using this more complete picture, Insight applies both rule-based and probabilistic machine learning algorithms to identify potential quality issues.

Insight has been fully deployed to adjudicative staff at the appeals level since late 2017 and the hearings level since late 2018. Internal studies by SSA of Insight's effect on adjudication have found that its use is associated with improved work quality (e.g., improved rates of quality issue remediation during drafting, improved quality issue recognition on appeal, etc.) and more efficient case processing.

The foundations described in Section IV were essential to Insight's development and operational success. First, SSA's existing case processing system provided data that enabled Insight developers to target specific pools of historic hearing decisions to streamline the assembly of labeled training data for machine learning features. The data infrastructure also enabled quality checks that rely on access to "ground truth" outside of decisional text to function. For example, a quality check of whether an age-related regulation cited in the decision is in fact applicable to the claimant requires the claimant's date of birth as stored in the case processing system.

Second, policies that rigidly structure the findings and content of disability decisions have been essential to the success of Insight's NLP and logic. For instance, the sequential evaluation process spelled out in Section II explicitly defines the findings that must appear in a decision and their sequence. This structure is manifested in decision templates used by staff (e.g., the template reliably outputs a "Step 1" finding prompt followed by a "Step 2" prompt, etc.). The predictability created by these policies greatly simplified the development of high precision information extraction.

Third, personnel flexibilities championed by OAO leadership enabled an adjudicative staffer to "pitch" Insight to SSA leadership and then have the flexibility outside regular legal duties to pursue its full development and release.

Finally, and critically, SSA ultimately invested millions of dollars over multiple years to transform Insight from essentially an "under the desk" proof of concept to a much more full-featured, enterprise-class system capable of supporting thousands of users. This enabled Insight to access highly skilled software development staff and contractors, as well as secure additional time from numerous highly knowledgeable business staff to help further guide the project and complete necessary data annotation tasks.

While SSA reports that Insight has improved quality and productivity, formal evaluations of the impact of the Insight system on accuracy and remand rates have been limited (Ames et al., 2020; Office of the Inspector General, Social Security Administration, 2019). There

is great potential for more rigorous evaluation and harnessing of more recent advances in AI. Many Insight quality flags, for instance, rely on more simple forms of machine learning and do not yet take advantage of the most important developments in deep learning with natural language processing.

Lessons

The SSA case study illustrates broad organizational and personnel challenges with AI innovation in the public sector. While OAO maneuvered around these bureaucratic impediments, AI innovation will require leveling these barriers more systematically. We hence spell out more general lessons to foster an improved ecosystem for AI innovation, accountability, and governance in the public sector.

Leadership support and blended expertise

A chief lesson from SSA is the critical role played by leadership at OAO to drive forward agency capacity to learn from its own data. Many of the key moves—capturing data, formalizing policy into an adjudicatory decision tree, leveraging individuals with blended expertise, making the strategic choice to invest in early analytics projects—would not have occurred in the absence of strategic leadership at the top.

SSA's experience also shows the deep value of "nexus" resources—those with both business and technical expertise—in driving AI innovation. First, nexus resources accelerate the speed of AI innovation. Leveraging their deep understanding of business operations, they can rapidly evaluate the potential value of an innovation along with its policy and cultural acceptability. Leveraging their technical expertise, they can often take major steps toward building and prototyping the innovation. By breaking the need to coordinate to devise, build, and validate, nexus resources tighten iterative development cycles, resulting in projects that fail or succeed much faster.

Second, nexus resources can increase the likelihood innovations will succeed. At federal agencies like SSA, significant funding for an innovation project requires a successful presentation of a well-researched business case to an investment review board. However, many ideas for innovation are complex to substantiate. While large organizations often have technical teams with the remit to pursue test cases, they are generally reliant on contractors or staff who are unfamiliar with the day-to-day business functions to which the innovation relates. The result is a knowledge gap that can delay or even stifle the substantiation of AI innovations. Nexus resources can bridge this gap by using their blended in-house expertise to brainstorm across functional teams, discover operational insights, and immediately build out a prototype.

Yet organizational structure often stands in the way. Many federal government positions are designed around specific sets of skills, with career tracks and promotions driven largely by work activity within those set areas. Blended expertise fits uncomfortably within these hired duties.

Leaders of organizations interested in AI innovation should consider how they can proactively structure their human resources to secure and foster nexus resources. First, more open-ended duties in position descriptions could provide the flexibility needed to leverage staff with nexus resources by retasking them to technical tasks under those duties. Getting position descriptions right will be critical for AI innovation (Engstrom et al., 2020; National Security Commission on Artificial Intelligence, 2021). Second, many agencies have drawn on partnerships with academic institutions—through vehicles like the Intergovernmental Personnel Act or Cooperative Research and Development Agreements—to bring in technical and nexus resources (Engstrom et al., 2020). One of the core challenges lies in the fact that AI innovation is occurring at a furious pace, and such partnerships, sabbaticals, and exchanges can leverage the core AI talent at research universities and the domain expertise of agencies.

The value of operational data

Data fuels AI innovation. SSA's Insight software is proof of the significant value earned from data describing the Disability Program's workflows and decision making generated in applications such as the eCMS. Even this data is only a foothold on a longer climb. If decision-facing AI innovations like Insight are ever to meet or exceed the accuracy and breadth of human counterparts, they need access to the same level of information as them. Moreover, robust data on workflows and decision making is often necessary to meaningfully achieve several principles of AI governance under Executive Order 13,690, such as accuracy, effectiveness, and traceability. It is difficult to evaluate if an AI system outperforms existing processes or is "accurate" without robust data cataloguing the operations it targets.

To achieve this, organizations can take steps to digitize operational activity and decision making through systems akin to SSA's eCMS. They also should strongly consider the passive collection of data, such as logging what is done and for how long in public and staff websites and other software. Passive data collection is highly cost effective because it requires no direct action to generate. Passive data also enables operational analyses—including analyses of the effectiveness of AI innovations[4]—whose probative strength and granularity are not possible without it. Indeed, leading private companies not only passively collect numerous forms of data, but they also act on it in real time.[5] While highly valuable, organizations should be cognizant that its collection is likely to raise ethical and even legal questions for those from whom it collects.[6] For example, staff-facing efforts could be introduced with commitments from management that the data would not be used for performance evaluations.

Rich operational data may be essential to identifying roadblocks to reaping value from AI innovations. For many AI applications whose outputs are engaged with by human staff, the AI's value will turn on whether it supports a task space that has a broadly agreed-upon structure and meaning by its users. For example, AI tools that classify animals in photos would have little global appeal if the rules of what constituted a cat versus a dog were subject to the whims of each user. Rich operational data enables organizations to meaningfully and efficiently evaluate how consistently staff perceive a task before investing in an AI application to support it.

For example, in SSA's case, data collected from its eCMS and case analysis systems enabled leaders to identify operational inconsistencies and target them through policy clarifications and training, ultimately improving adjudicative consistency. This process was incredibly valuable to Insight's acceptance by users, as they were more likely to share and agree with Insight's applications of policy in its quality checks. Organizations should be ready to consider pursuing some level of "upstream" policy and process reforms to promote consistency in a business task as needed before injecting supportive AI into it.

Test early and often

SSA's experience shows that testing AI innovations through pilots or limited releases can be a valuable means to evaluate the value of the innovation and build organizational buy-in. For example, SSA released Insight to a subset of OAO adjudicative staff in early 2017 for voluntary use in their cases. SSA then studied how Insight impacted OAO operations, which in part showed that Insight use had no deleterious impact on case processing efficiency—a concern prior to its release. SSA also surveyed staff who tried Insight, and a large majority of respondents indicated they found Insight feedback to be accurate and easy to interpret. These results were critical to OAO's decision to expand Insight's release and provided objective support for SSA leadership outside OAO to increase the Insight project's funding to expand its features and scale to the hearing level. These sorts of evaluations will yield valuable insights into how AI systems operate in practice, including the potential for unanticipated effects (Engstrom & Ho, 2020).

Although there have been some efforts to define the word "pilot," we caution that the term should not automatically trigger layers upon layers of agency or congressional review. Review should be calibrated based on the risk posed. AI systems to augment existing quality improvement programs, for instance, may be precisely the kind of internal organizational decisions for which management flexibility is warranted.

Continuous iteration and evaluation

The SSA experience also demonstrates that public sector AI innovation is a process, not a product, requiring continuous analysis, evaluation, and iteration.

Consider the use of supervised learning. Recent advances in deep learning are particularly appealing because logical AI systems—models based on hard-coded conditional logic—are challenging to scale, given the thousands of fact scenarios evaluated under complex and often vague policy rules. Supervised learning, instead, requires only labeled examples of a targeted decision (e.g., "disabled" vs. "not disabled").

Yet precisely because of the historical challenges in decisional accuracy, securing enough high-quality labels to train models will be an ongoing process. Many federal agencies may be facing an explosion in their capacity to collect and analyze operational data, due to technologies such as cloud infrastructure, handwriting recognition, and speech-to-text. Through these advances, agencies may be beginning to learn far more about the quality and consistency of their past decision making. For example, with the advancement of technologies that can extract detailed data from claimant medical records, SSA may

identify disparities in outcomes among highly factually similar historic claims, such as may be caused by inconsistent heuristics used by staff. Any such anomalies in past actions may equate to a ceiling for the performance of supervised learning models based on them, and agencies may not be comfortable with that ceiling.

If systemic errors, biases, or inconsistencies are exposed, one might be tempted to simply filter them out. However, a more viable path may be to use what is uncovered to improve *future* decision making through a *combination* of analysis, early-stage AI, and human judgment. For example, analysis could reveal inconsistencies among staff in applying a particular regulation. Targeted training or human-in-the-loop AI features could improve consistency in this area. With a combination of targeted improvements, the quality and consistency of human decisions can often be improved and thus offer a better position from which to train a supervised AI system. This approach acknowledges the clear reality that human decision making may include flaws and instances of bias, as well as many virtuous attributes. For example, SSA's disability adjudicators may make observable mistakes, but they also make thousands of discrete decisions that deftly navigate complex policies and medical fact settings. After a period of AI-augmented, well-analyzed human performance, agencies may finally have a series of historic decisions that are of sufficient quality to train a responsible decision-*making* AI system.

However they approach their historic actions, organizations reading SSA's example and seeing AI as a one-time panacea are cautioned to consider it as part of continuous improvement. Even if an AI-based tool does no worse than an existing human baseline, such AI-based tools may themselves generate the impetus for further performance improvement. Indeed, the logic of *Mathews v. Eldridge*—which would point to the lower administrative burden of reducing error with AI-based tools—may demand it.

Development and deployment ecosystem

Organizations often develop software using multiple development environments. Prior to release to a "production" environment (real world operations), a "validation" environment is often used to test features using "mock" data designed to be structurally equivalent to production data without corresponding to any real-world entity. However, this paradigm can impose several significant disadvantages particularly harmful to data hungry AI innovations. First, the creation of "mock" data can be a slow, request-based process whose size does not compare to production. Second, the "mock" data may not be complete or faithfully represent complexities in the underlying data. For example, agencies may find it difficult to meaningfully mock unstructured data such as legal motions or customer service transcripts.

This infrastructure is not optimal for the rapid, realistic, and massive scale experimentation and testing of new AI features. AI development often requires significant computational resources, the use of various open-source software packages, and the full feature space, enabling better modeling and error analysis. Agencies should pursue infrastructure that will allow teams to safely but easily prototype and test new innovations at scale against real-world production data. For instance, agencies could establish a "data warehouse" containing a replica of production system data that development teams could ingest from to experiment and validate new AI features.

Additionally, organizations should consider how their broader software ecosystem enables and integrates with AI use cases. AI innovations do not work in isolation; their outputs must be presented to users or effectuated within existing systems to generate value integration into existing systems. But existing systems often are not designed with integration in mind, forcing AI teams to spend significant time and money on "plumbing" issues unrelated to their core objectives. To help AI innovations scale without this structural friction, organizations should consider requiring core systems to be engineered to facilitate rapid extension. For example, a workload system could offer a secure API endpoint that would enable a calling system to programmatically make changes to individual work items or push custom content to a portion of the user interface for review by users.

Conclusion

An evolution toward integrating AI into the recurring workloads of large organizations may be inevitable. SSA has nearly 60,000 employees, with administrative costs exceeding $6 billion annually.[7] Every merits decision in SSA's Disability Program is made by human staff with naturally varying levels of expertise and perceptions of often vague policies. Insight's release has proven that AI partnered with human staff can help counter these drawbacks to generate significant improvements in performance over humans alone.

SSA's example teaches us that for any AI system to achieve good governance, organizational reforms to enable invention, reflection, and assessment will be critical.

Notes

1. Affiliations are for identification only. This research does not represent any official views or opinions of the SSA. This article draws on prior work by two of the coauthors (Bajandas & Ray, 2018; Engstrom et al., 2020; Engstrom & Ho, 2020; Ray & Lubbers, 2014; Ray & Sklar, 2019). Ray was an Administrative Appeals Judge and Deputy Executive Director of the Office of Appellate Operations at SSA and spearheaded numerous initiatives documented herein. Glaze is a Program Analyst and the Creator and Product Owner of the Insight Software at SSA. We thank Nikita Aggarwal, Justin Bullock, Johannes Himmelreich, Sonoo Thadaney Israni, Michael Matheny, and Padmashree Gehl Sampath for useful comments. Ray and Ho were consultants on Administrative Conference of the United States, Recommendation 2021-10, Quality Assurance Systems in Agency Adjudication, https://www.acus.gov/research-projects/quality-assurance-systems-agency-adjudication (recommending the use of data analytics and machine learning for quality assurance).
2. Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, Exec. Order No. 13,960, 85 Fed. Reg. 78,939 (2020).
3. Co-author Kurt Glaze, an SSA attorney with an interest in computer science, devised and pitched Insight to OAO in 2015.
4. For example, if an organization implemented an AI system to automatically update customer mailing addresses based on the free text within various inbound mail, they may well want to benchmark how human staff perform that task. Without scaled passive data recording, benchmarking would likely consist of observing a small number of examples

or simply asking staff about the task, both of which may miss critical insights into as-is performance.

5. For example, Amazon introduced AI camera systems into their delivery truck fleet to monitor drivers for potential traffic violations and distraction, e.g., drowsiness. Drivers must sign consent forms acknowledging this monitoring. Amazon reports that the AI technology has contributed to massive improvements in driver performance: "[A]ccidents decreased 48 percent, stop sign violations decreased 20 percent, driving without a seatbelt decreased 60 percent, and distracted driving decreased 45 percent" https://www.theverge.com/2021/3/24/22347945/amazon-delivery-drivers-ai-surveillance-cameras-vans-consent-form.
6. For example, the European Union's General Data Protection Regulation gives individuals from whom businesses collect personal information (including web analytics data) numerous privacy rights. See https://gdpr.eu/tag/chapter-3/.
7. https://www.ssa.gov/oact/STATS/admin.html.

References

Ames, D., Handan-Nader, C., Ho, D. E., & Marcus, D. (2020). Due process and mass adjudication: Crisis and reform. *Stan. L. Rev*. *72*, 1.

Bajandas, F. F., & Ray, G. K. (2018). *Implementation and Use of Electronic Case Management Systems in Federal Agency Adjudication*. https://www.acus.gov/report/final-report-implementation-and-use-electronic-case-management-systems-federal-agency

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and c ommunication technology is transforming administrative discretion and constitutional control. *Public Administration Review 62* (2), 174–184.

Bullock, J., Young, M. M., & Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity 25* (4), 491–506. https://doi.org/10.3233/IP-200223.

Busch, P. A., & Henriksen, H. Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity: The International Journal of Government & Democracy in the Information Age 23* (1), 3–28. https://doi.org/10.3233/IP-170050.

Casado, M., & Bornstein, M. (2020). The new business of AI (and how it's different from traditional software). *Andreesen Horowitz*, 2019–2020. https://future.a16z.com/new-business-ai-different-traditional-software.

Cinar, E., Trott, P., & Simms, C. (2019). A systematic review of barriers to public sector innovation process. *Public Management Review 21* (2), 264–290.

Criado, J. I., Valero, J., Villodre, J., Giest, S., & Grimmelikhuijsen, S. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity: The International Journal of Government & Democracy in the Information Age 25* (4), 449–470. https://doi.org/10.3233/IP-200260.

Edwards, H. T., & Livermore, M. A. (2008). Pitfalls of empirical studies that attempt to understand the factors affecting appellate decisionmaking. *Duke LJ 58*, 1895–1989.

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic accountability in the administrative state. *Yale J. on Reg. 37*, 800.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *Report to the Administrative Conference of the United States*. https://www.acus.gov/report/government-algorithm-artificial-intelligence-federal-administrative-agencies

Gelbach, J. B., & Marcus, D. (2017). Rethinking judicial review of high volume agency adjudication. *Tex. L. Rev*. *96*, 1097.

Gelbach, J. B., & Marcus, D. (2016). A study of social security disability litigation in the federal courts. *Final Report to the Administrative Conference of the United States*, 16–23. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2821861.

Government Accountability Office. (2020). *Social Security Disability: Information on Wait Times, Bankruptcies, and Deaths among Applicants Who Appealed Benefit Denials* (GAO-20-641R). GAO, Washington, DC. https://www.gao.gov/products/gao-20-641r.

Ho, D. E. (2017). Does peer review work: An experiment of experimentalism. *Stanford Law Review 69*, 1–120.

Mashaw, J. L. (1973). Management side of due process: Some theoretical and litigation notes on the assurance of accuracy fairness and timeliness in the adjudication of social welfare claims. *Cornell L. Rev*. *59*, 772.

Mashaw, J. L. (1985). *Bureaucratic justice: Managing social security disability claims*. Yale University Press.

Mashaw, J. L., Goetz, C. J., & Carrow, M. M. (1978). *Social security hearings and appeals: A study of the Social Security Administration hearing system*. Lexington Books.

Moldogaziev, T. T., & Resh, W. G. (2016). A systems theory approach to innovation implementation: Why organizational location matters. *Journal of Public Administration Research and Theory 26* (4), 677–692.

National Security Commission on Artificial Intelligence. (2021). *Final Report*. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

Office of the Inspector General, Social Security Administration. (2019). *The Social Security Administration's Use of Insight Software to Identify Potential Anomalies in Hearing Decisions* (A-12-18-50353). https://oig.ssa.gov/audit-reports/2019-04-22-audits-and-investigations-audit-reports-A-12-18-50353/

Ray, G. K., & Lubbers, J. S. (2014). A government success story: How data analysis by the Social Security Appeals Council (with a push from the Administrative Conference of the United States) is transforming social security disability adjudication. *Geo. Wash. L. Rev*. *83*, 1575.

Ray, G. K., & Sklar, G. (2019). An operational approach to eliminating backlogs in the Social Security Disability Program. *McCrery-Pomeroy SSDI Solutions Initiative*. https://www.crfb.org/sites/default/files/An_Operational_Approach_to_Eliminating_Backlogs_in_the_Social_Security_Disability_Program.pdf

Rubenstein, D. S. (2021). Acquiring Ethical AI. *Florida Law Review 73*, 747–819.

Sayer, P. (2016, October 24). Not robocop, but robojudge? A.I. learns to rule in human rights cases. *Computerworld*.

Tashea, J. (2019, June 7). France bans publishing of judicial analytics and prompts criminal penalty. *ABA Journal*. https://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics

Tushnet, M. (1980). Post-realist legal scholarship. *Wis. L. Rev*. 1383.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance 2* (4), 301–313.

CHAPTER 39

WATCHING THE WATCHTOWER

A Surveillance AI Analysis and Framework

STEPHEN CAINES

Introduction

Governmental surveillance has radically expanded in scope and sophistication through the development of Surveillance AI (SAI). SAI is defined as any instance where a governmental surveillance activity is supplemented or facilitated with the use of artificial intelligence (AI). Technological innovations like facial recognition that were previously reserved for only the largest and most resourced levels of governments are now accessible to even the smallest jurisdictions and even individual citizens. This transition has been ushered in by the removal of impediments in education, computing power, talent, and data. The global appetite for SAI is continually growing: 75 of 176 surveyed countries are actively using some form of AI for surveillance (Nouri, 2020). A wide array of use cases currently exist—AI predicts potential terrorists on flight manifests, controls autonomous drones that surveil from above (using computer vision), and screens conversations for key words to assist law enforcement (using natural language processing). Unfortunately, the development of SAI is occurring faster than the oversight, safeguards, and regulations can follow. Thus, SAI is often prematurely deployed despite incompatible stated goals, which can result in harms as minimal as small privacy invasions and as large as human rights violations.

Given the highly controversial nature of surveillance, as well as the complex historical, technical, sociological, privacy, and ethical underpinnings, it can be difficult to discuss the individual and collective merits and impacts of these deployments. Too often modern discourse centers on a binary good or evil AI assessment that fails to adequately address the nuances of each situation. Further, many fail to recognize the dynamic and fast-paced rate at which these technologies evolve. The burgeoning field of AI Ethics has made great strides in recent decades but due to deficits in resourcing, talent pools, and implementation it has only barely begun to deliver on its potential. Additionally, industry leaders such as Google have been fraught with allegations that they are deliberately and intentionally

silencing AI Ethicists who are charged to ring the proverbial alarm about issues on the horizon (Schiffer, 2020). The discipline currently succeeds at identifying relevant potential problems that may arise or persist but frequently grapples with gaining consensus for best practices, surveying the entire landscape, and playing a role in all parts of a tool's development. The issue is further complicated by government users of technology having limited in-house staff with formal training in AI Ethics. As a result, audit checks occur only at deployment, if at all, and not during earlier stages of ideation and development. These ground truths require the responsibility for ethical deployment to be diffused from solely AI Ethics experts to a wider population. For any SAI stakeholder—buyer, vendor, user, or subject—one must have a framework to assess how the technology has been ethically created, changed in complexity as well as in use, and subsequently ask whether its continued use can be justified.

Two useful concepts capture the implications of widespread governmental SAI use: mission creep and function creep. These concepts are often used interchangeably, but this chapter will use them independently. In a basic sense, mission creep is the same technology being used for new purposes or on new subjects, while function creep is modified or more sophisticated technology being applied to similar subjects or networks.

Mission creep assesses the continuity between the proposed reason for implementation and the ground truth. Mission creep arises when there is a disconnect between stated goals and reality. The critical issue here is not that the technology is solely being used for a new purpose, but rather that the original implementation occurred under a certain set of conscious assumptions and agreements, whereas this new use case has not received equal scrutiny. Had another balancing test of costs and benefits occurred, the novel use case may not have been approved for use. Moreover, mission creep can also be defined as the use of existing technology in new and often unassessed scenarios that produce unanticipated consequences. Mission creep in SAI becomes a nascent issue in three scenarios: (1) when deploying existing and approved technology on new subjects or in new scenarios; (2) when an existing technology begins to derive novel and unintended inferences or consequences; and (3) when organizational priorities and administrative ideals undergo a significant change or shift, such that SAI use is used in a new manner.

An example of mission creep is the use of Automated License Plate Readers (ALPRs) in the United States, which were originally introduced as a means to find kidnapped children and recover stolen vehicles. These original use cases were not very controversial as the public finds those types of crimes less sympathetic and this particular tradeoff of having their movements monitored through automated systems is a worthwhile trade. But not long after, the US Immigration and Customs Enforcement agency (ICE) used ALPRs to determine the whereabouts of undocumented persons (Zubair, 2019). In addition to this being a much more controversial issue—numerous jurisdictions have declared themselves as sanctuary cities who do not voluntarily collaborate with ICE—there is another issue at play. Had the initial implementation of ALPR technology included a clear use case of immigration, the subsequent discussion and decision of whether this technology should be used would likely have been different.

Another example of mission creep in technology is the Singaporean government's decision to use contact tracing information—originally intended to protect public health—for criminal investigations (Sato, 2021). To make matters worse, the government did not notify its citizens that data would be used for that purpose.

Function creep is distinguished as the tendency for technologies themselves to grow and shift in complexity, sophistication, and capability without sufficient oversight. Whereas mission creep often occurs at a slower pace and in a more linear fashion, function creep can be imagined as the branching of a tree or a river with diverging forks. The growth or shift in complexity, sophistication, and capability requires talent and makes function creep primarily an issue that occurs with producers and vendors of SAI technology instead of with its users.

A key example of function creep are municipal CCTV networks managed by the local police jurisdiction—say, the NYPD in New York—when this network is connected to a cloud platform to provide services such as real-time facial recognition. What was a passive sensor network that was too large to reasonably be staffed with humans is now transformed into a real-time monitoring tool that could track the movements of millions of people a year in an indiscriminate and automated manner. The key issue here is that in order to achieve this functionality, a "silent upgrade" is required; that is, not just an improvement of technical function (in this case using the camera feeds to fuel a deep learning algorithm), but rather a significant improvement of function without public notice or comment, often for fear of resistance. Had the parties subject to these enhanced cameras been made aware, they may have delayed or prevented the process through civil resistance like petitions and open meetings. New York is currently facing widespread criticism for spending a "slush fund" of $159 million over 14 years to purchase and maintain SAI and other tools without public oversight (Holt, 2021). Function creep is not the general tendency for technology to improve. Instead, it's the deliberately silent or obscured improvement that evades detection or accountability.

This perfect storm of SAI technical evolution and loose regulatory structure has given rise to thousands of surveillance deployments across the globe, some of which produce existing human rights violations, diminution of general and personal privacy, and, at times, an exacerbation of the marginalization of society's most vulnerable.

To assist in assessing the technical, legal, and ethical integrity of different deployments, this chapter will attempt to arm readers with two skills while providing a broad survey of the current SAI landscape: First, developing the ability to independently assess different SAI deployments for technical soundness, presence of mitigating safeguards, and a basic ethics assessment; and second, using notions of mission creep and function creep to identify and prevent systematic and frequent negative impacts of SAI use.

This chapter will discuss the distinguishing factors or SAI, the surveillance technology hierarchy, the Seven Axes of Surveillance, trends in SAI, mitigating SAI harms, and a consideration of the future. Cumulatively, these sections will provide a framework for assessing both contemporary and future global SAI government Deployments.

What Is SAI: Three Distinguishing Factors

SAI has three defining characteristics: (1) a reliance on big data; (2) private–public partnerships; and (3) alignment with existing political structures. While the first two characteristics can be found in almost any field of AI application, the third adds an important layer of complexity.

Reliance on big data

If the AI at the center of many surveillance deployments is analogous to an engine, big data is undoubtedly the gas that fuels it. The three Vs—Volume, Velocity, and Variety—summarize key aspects of big data. AI systems typically rely on large troves of data to train the initial algorithm and to produce queries or algorithm outputs through models. Highly functioning systems typically require data that is structured (meaning the presence of features like naming conventions and common modes of organization) and labeled (the process of adding an informative layer of information to raw data). The data can take the form of images of people or objects; biometric readings, like a face template; audio files or frequencies; scientific readings or measurements; personally identifiable information, such as addresses or social security numbers; or other information. In addition to the format of the data, data integrity is equally, if not more, important. Accurate assumptions, correlations, and readings are vital as the common paradigm of "garbage in; garbage out" suggests. The world's most sophisticated algorithms will fail if provided with irrelevant or incorrect information. These three features are often evaluated and implemented in the "data cleaning" aspect of the development pipeline. Data integrity is at greatest risk when there is insufficient data scrutiny or shortcuts are taken in the cleaning stage.

The origin of big data is often among the most controversial parts of generating a Surveillance AI system due to a pervasive lack of consent across different use cases. Early AI surveillance systems relied on consent because the conditions required for clean data collection required the subjects to affirmatively grant their biometrics or information to be collected. Collection was typically dependent on them coming to a specific location, like a laboratory or a set, to have readings taken with specific machines. In the context of facial recognition, many of the earliest datasets were consenting college students being recorded on university campuses. Later datasets were created by placing covert cameras in highly trafficked areas like coffeeshops to record images without subjects' consent. In modern times, some facial recognition datasets are created by data-scraping social media—again, without consent—which, to make matters worse, unlike the coffee shop example, takes not just a sole glimpse of someone but at times hundreds of pictures of a person over the years. This change was ushered in with the sharing and brokerage of datasets (an over $225 billion industry), and the advent of the Internet of Things (Knowledge Sourcing, 2021).

Big data is not merely sold but is also shared, and government SAI systems can be created with internal or external sharing of data (Georgetown Law, n.d.). A 2016 Georgetown report revealed that one in two Americans are in a facial recognition database. Department of Motor Vehicle (DMV) databases have been used for a variety of purposes, such as training biometric identification systems like facial recognition and to run queries for ICE (Harwell, 2019). It is key to highlight that when individuals apply for a driver's license there is rarely, if ever, notice that the information provided may be used against them in such contexts. Departments are incentivized to share access because the network effect works to improve the results of the system as more databases are consolidated. This practice is often most harmful in states like California which provide driver's licenses to immigrants regardless of citizenship status and information is shared with a variety of agencies (National Immigration Law Center, 2016). Data sharing also occurs across borders so your information may be in the hands of countries you've never been to. The Common Identity Repository (CIR), for example, is a massive biometric database currently under construction, consisting of

European citizens, visitors on a visa, and third-country nationals (Pivcevic, 2020). The CIR has the potential to include 350 million people because it combines six existing databases in the EU (Keck, 2019).

Aside from sharing, the data brokerage industry is valued at over $200 billion (Whittaker, 2020). Much of the data capture occurs by private industry and is sold to government contractors with very few details provided to the public. In November 2020, the location data company X-mode was found to be selling data from the Muslim prayer app "Muslim Pro" to the US military (Cox, 2020). The app has been downloaded by over 98 million people worldwide, and the incident was highly controversial because Muslims in America have been historically over-surveilled. Additionally, there are national security concerns that other countries may be using applications like TikTok (a social media company created by the Chinese company Bytedance) to surveil Americans (Brown, 2020). These fears have manifested in restrictions for government employees and devices, as well as considerations of an outright ban from new downloads. Finally, the Internet of Things (IoT) has impacted this space by permitting legacy systems of technology like CCTV to be upgraded with complex SAI, such as face and object detection—a case of function creep (Privacy International, n.d.). The increasing availability and decreasing price of cloud computing have been driving this surveillance renaissance. These factors have all culminated in high demand and supply of SAI.

Big data directly feeds into function creep as greater nuances, distinctions, and possibilities for groups of individuals continue to arise almost daily. It then becomes a system where output is no longer limited by fuel but by the design of the engine. The data brokerage industry has created a significant incentive to capture and store interactions first, and analyze them second. The result is that data has an actualized value on the day of capture as well as a potential value for future, currently unknown, developments. Big data can also impact mission creep because if an existing system is fed new information, such as new individuals, in a similar context, an agency could expect similar results and utility.

Private–public partnerships

The private and public industries have a strong symbiotic relationship in developing surveillance technology. The government's role is not solely limited to consumers, they also serve as regulators and producers of the technology. While the government has started many technological initiatives and projects like the modern internet, the primary entity driving SAI specific innovation in terms of capital and talent is the private industry (Manjoo, 2012). The private sector enjoys an abundance of technical expertise, the benefit of agile development, the ability to raise massive amounts of capital, and the flexibility of experimenting with organizational structure (Brown & Ezike, 2020). They are often able to leverage their freedom to develop technological solutions that laws and regulations are decades behind on potentially regulating. Government has access to highly sensitive, valuable, and high quality data, the authority to deploy systems for general welfare and law enforcement purposes, and is not subject to certain restrictions regarding privacy faced by the private sector. Currently, many government SAI systems are built on a contract basis or purchased through a subscription model with the assistance of a Private Vendor. This is a direct result of the decrease of in-house technical development expertise in the government. While this has likely increased the pool of technology being developed and marketed, it has come at a

price. The private vendors that create the technology may lack context that may aid product development and similarly often lack incentives to continue improvement or agile development after the contract term ends (Greene, 2018). Additionally, lawmakers may avoid traditional legislative means and simply purchase technology to reach their goals, a process known as "policy making by procurement" (Crump, 2016).

Universities play a vital role in the development of SAI as the perfect crucible for development (Owen, 2019). Aided by their tax classification as not-for-profit institutions, a revolving door of young technical talent, and freedom to research with little oversight or constraints, they have been the starting point for several SAI deployments. In addition to operating as the starting point for innovation, technology licensing and transfer operations often serve as the first step of commercialization for SAI. In this process SAI developed by university students, staff, or affiliates undergoes a technical audit and sale or licensing to an interested party for a definite or indefinite time period (Bedoya, et al., 2016). While universities also license intellectual property, such as original music and other research, SAI licensing has specifically undergone additional scrutiny in recent years due to the implications. Palantir, a data mining brainchild of Stanford, has specifically drawn ire for their work with the Department of Health and Human Services (HHS) and ICE (Steinberger, 2020). The minimal restrictions on technology licensing agreements and incentive structures frequently prioritize profit over true oversight.

It is vital to note that arrangements between the private and public industries are not always formally established. One may assume that the use of such powerful technology naturally comes with a formal agency contract followed by a public vote from an oversight board. Oversight provides a chance for notice, assessment, input, and monitoring. All four of these elements provide a key role in deploying and mitigating the negative impacts of such disruptive technology. It is key to note private–public relationships can also exist outside of formal agreements or contracts. Some of the most ethically gray areas involve law enforcement being given access to private civilian or business networks, like the Ring doorbell (Lyons, 2021). Access can be granted on an individual basis or instance, such as when a crime is committed on or near your property and you grant law enforcement access to your device or network. Alternatively, groups and individuals can grant general access to law enforcement to passively monitor or gain unfettered access. Groups that pool resources to purchase SAI often make democratic decisions about the system's utility and purpose. A common model of the latter has been implemented by the Community Benefit Districts, a partnership between commercial and mixed use property owners (SOMA West, n.d.). In San Francisco, communities pool resources to purchase surveillance technology. The issue that arises, though, is that the technology cost is often subsidized by groups or individuals with potential conflicts of interests, such as in San Francisco (Bowles, 2020). Additionally, some of the standards of who can access what and when are often kept deliberately vague or obscured from public view. It is key to define the system at the center of focus as private, public, or hybrid to understand the authority, obligations, and duties of the system's users and benefactors.

Correlation with political governance

While it can be difficult to make normative statements around SAI use, there are clear trends between the style of political governance and geopolitical pressures. The first key

observation is that while almost all modern developed nations have at least experimented and at times used SAI, authoritarian nations are significantly more likely to build and deploy that technology. China is the largest exporter of SAI as four of their companies—Huawei, Hikvision, Dahua, and ZTE—supply technology to 63 countries. China not only exports the technology but widely uses it internally on its own citizens. A number of cities like Shanghai have automated the detection of minor crimes, like littering, and use facial recognition and cell phone location to automatically bill alleged violations (Dai, 2019). The systems are so advanced that a man driving in Jinan was ticketed for scratching his face in a traffic jam when an SAI believed he was on his phone (News from Elsewhere, 2019). This information is often used to create a social credit score which has a number of implications on housing, employment, and quality of life. While the US also has consequences for the violation of certain laws, the extent to which the detection and prosecution of these crimes are automated is far less. It is also key to note that the most vulnerable are often the most surveilled in any nation. The implications can be as minor as a higher density of surveillance cameras in impoverished areas of major American cities or the ethnic tracking and reeducation of the Uighur Muslims in China. Authoritarian regimes are more likely than democratic regimes to gravitate towards and use SAI solutions. As mentioned previously, however, all nations engage in some amount of digital repression which can be direct applications of technology or failing to account for natural differences in populations.

The Surveillance Technology Hierarchy

Surveillance

It is key to distinguish which subgroup of surveillance technology is being discussed because they each have their unique amount of considerations. Traditional surveillance was analog in nature and relied heavily on the ability of human resources to drive the generation of useful information. In the US, the story of surveillance is one largely defined by drawing lines between citizens and second-class citizens or slaves. Some of the earliest proto-biometric databases were plantation ledger books to keep information and notes about slaves (Chao et al., 2019). These ledger books assisted slave owners and the government in distinguishing free and full citizens from their enslaved compatriots. Importantly, only those in the lowest socioeconomic ladder were subjected to this form of surveillance.

What makes government surveillance distinct from private surveillance is the ability to compel subjects to participate and help the identification, lessening the burden or energy expended by the surveyor. In the eighteenth century, New York required all enslaved persons to carry a lantern at night when not in the presence of a white person for easy identification (International Dark-Sky Association, 2021). Tactics such as these not only create an obligation on the surveyed but also serve as a stark hierarchical reminder. A modern example of this would be as recent as 2014, when the then New York mayor had floodlights pointed at public housing projects in the name of deterring crime (Gellman & Adler-Bell, 2017).

Whether by race, socioeconomic status, religion, or orientation, governments have relied on surveillance as a means of maintaining class structures and power. Just as the lanterns were borne by the most oppressed, the poorest individuals currently face a disproportionate amount of surveillance relevant to more affluent communities.

Surveillance technology

Instead of relying purely on human resources to collect information, a new era of surveillance uses technology—sensors and computing, data collection, and analysis—to collect and interpret information. The technology allows using dragnet searches instead of looking for specific subjects. Subjects are no longer directly chosen for surveillance; instead, large troves of data are collected and data analysis homes in on persons of interest in these data.

One example of this development is the addition of a data analysis technique to locate grow houses for marijuana (Durbin, 2019). Previously houses and locations would be revealed through other evidence or informants, a warrant would be sought for that specific property, and executed if granted. Agents could then likely use a thermal camera on that specific location (Campbell, 2021). By contrast, in the data-driven dragnet model of surveillance, police survey an entire city and create a short list for investigation by analyzing power usage and marking suspicious locations.

The term "Surveillance Technology" is a widely inclusive term with several key distinctions and subgroups. Varieties of surveillance technology exist in telecommunications, video, aerial, social media, and audio spaces. The term also includes other surveillance tools that do not use AI. Examples of Surveillance Technology that do not fall within the scope of surveillance AI include devices like Stingrays—a device that simulates cell towers to determine the location of cell phone data. Although SAI systems may be more complex, from a technical standpoint this does not inherently mean that they are more efficient or equitable.

Surveillance AI

It is important to distinguish between Surveillance AI and other forms of Surveillance Technology as the methods and practice of governance among the different groups of technology can differ greatly.

SAI is distinguished from general surveillance technology as it does rely on AI. One of the fastest growing areas is Biometric Surveillance which observes and uses physiological characteristics of individuals to identify them. A robust biometric surveillance system must rely on an immutable feature; that is, a feature would be resistant to change. For example, a system that depends on hair color would not be as reliable or effective as a retina scan because hair color can be easily changed. While certain forms of biometric surveillance, such as facial recognition, are widely known, new forms are rapidly being discovered and developed. A few of the most novel applications of biometric surveillance include Vein Matching, Ear Print Analysis, and Gait Detection. Given the need to create and store a biometric template of individuals for comparison, data breaches from biometric surveillance systems are particularly damaging because the information used to train the model can later be used to identify those same individuals in another context.

The layering effect

It is key to note that in modern SAI deployments the boundaries between the aforementioned categories are often blurred and combined. One source of information (e.g., a camera feed or a data set) can serve a number of different purposes in the SAI deployment pipeline. For example, a CCTV with audio and video capabilities placed in New York's Times Square could be the gateway that enables object detection visual algorithms to identify any potential weapons in a crowd, implement gait detection for specific individuals, and analyze crowd flow patterns to better delegate resources and personnel (Amnesty International, 2021a,b). The ambitious Domain Awareness Center, which was proposed in Oakland, California combines a number of technologies like cameras, gunshot detection systems, and license plate readers (Wheeler, 2016; Robertson, 2014).

The Seven Axes of Surveillance

In the following discussion, I provide a contextual framework that can be used to examine different surveillance use cases. The framework consists of Seven Axes: (1) The Surveyor, (2) The Surveyed, (3) Methods and Instrumentalities, (4) Passive/Real Time Monitoring, (5) Subsequent Effect, (6) Accountability, and (7) Transparency. The Seven Axes have been summarized in Figure 39.1.

Potential situations to use this framework include evaluating governmental use of SAI, implementing an SAI system internally for safety/worker productivity, and relying on SAI to determine insights or target potential clients or customers. Governmental use cases should feature special analysis on Subsequent Effects as well as Accountability and Transparency.

The surveyor

The term "surveyor" relates to the creator, user, purchaser, or beneficiary of the SAI system. Their identity often determines the structure, legality, and practices around use.

Axis Number	Axis Name
1	The Surveyor
2	The Surveyed
3	Methods & Instrumentalities
4	Passive/Real Time Monitoring
5	Subsequent Effect
6	Accountability
7	Transparency

FIGURE 39.1 The Seven Axes of Surveillance

In situations where the government is directly collecting the information and processing it, their practices are typically subject to relevant local, state, and federal law. Many forms of modern government surveillance practices have been implicitly or explicitly authorized in legislation, such as the post 9/11 USA/Patriot Act and the 2018 FISA (American Civil Liberties Union, 2021). The legal remedies in the US have been limited. The few causes of action resulting from the government's use for individual and class action plaintiffs often fail for reasons such as a defense of national security, failure to state a particularized harm, or State Secrets Privilege (American Civil Liberties Union, n.d.). To further complicate the matter, certain specialized courts like the Foreign Intelligence Surveillance Court have historically always ruled in favor of government surveillance without providing minimal explanation or justification. A landmark case, *Federal Bureau of Investigation v. Fazaga*, is currently being debated by the Supreme Court to determine whether unlawful government surveillance may be adjudicated under FISA instead of being dismissed under the States Secrets Privilege (Mansoor, 2021).

The surveyor—the owner, operator, or creator of an at-scale SAI system—is inherently operating at a considerable significant power differential with the subjects of the system. While basic systems can be created with some basic programming experience and online tutorials, access to information and professional talent to derive insights needed by municipalities tend to restrict the most effective surveillance to heavily funded and resourced individuals. Their actions are most strongly tied to mission creep and function creep because they generally have the power to mandate the project development and use. Recognizing this allows one to not only understand a system's current structure but also to address the trajectory of use.

The surveyor may or may not leave a paper trail of their purpose or acquisition of SAI systems. US government agencies have been minimizing such paper trails as they have been gifted SAI technology, been given a free trial of the technology, use the technology on an individual agent basis without formal departmental authorization, enter agreements that hide the real surveillance purpose (where the agency pays primarily for storing the data or receiving some routine maintenance or service), or give the vendor an economic stake in the fines or results generated by the system (Fassler, 2021). The incentive to not have a paper trail or provide notice of its use often relates to concerns of public image and perception, the burden of public records requests, and the likelihood of fiscal scrutiny. Mandating the creation of a paper trail of acquisition and use is vital to both internal agency policy as well relevant legislation ensuring the protection of the public's welfare.

When analyzing the surveyor, three questions should be asked: (1) Who built the technology? (2) Who funded the technology? and (3) Who has the authority to dictate the use or development of the technology?

The surveyed

The second axis of surveillance, the surveyed, can occur in an individual/particularized, mass collection, or targeted population fashion. Particularized collections typically rely on one piece or instance of data within a database, like traveler photos collected at the border. The most common form of mass collection is in the law enforcement and administrative agency context. For instance, the Internal Revenue Service has deployed AI to supplement

their search for individuals who have underpaid on their taxes by better allocating their auditing resources (RegLab, n.d.). But this surveillance is based on mass collection. The information is coming from their internal records of past confirmed underpaid cases to predict the traits of other likely plaintiffs. As such, everyone who filed taxes and is in their database would be surveyed. Finally, a prime example of targeted population surveillance is the case of Muslims following 9/11 when Muslims throughout the country were heavily surveilled by local, state, and federal authorities (Khan & Ramachandran, 2021).

By simply asking who is being surveyed, it can be hard to see the forest among the trees. The inquiry must not only stop at the individual designed to be targeted, but must also include the classes of people who may be incidentally exposed. For instance, when implementing an internal system to track employees, many people forget there are also maintenance, cleaning, clients, counsel, janitorial, and repair staff that also regularly enter the buildings. Determining who all is being surveyed can frequently result in an undercounting of individuals due to situations such as the one described. This is particularly urgent because surveillance tends to focus on political opponents, social outcasts, the unhoused, low income, and LGBTQ individuals, a trend discussed in detail later in this chapter, but also can implicate their family and people in their network (Georgetown Law, 2019). When analyzing who is surveyed, the question of "who all" is just as important as why.

Another key aspect of SAI systems is that they are often tuned with data from the surveyed at two critical points in the product life cycle: (1) Training/Development, and (2) Launch/End-User Operation. Tuning is the process of model optimization by changing hyperparameters. For instance, if the SAI deployment was monitoring a youth-based social media app to flag references of drug use, the original algorithm may be trained off of Reddit conversations, which is a more of a young adult/middle-aged platform. Tuning in this scenario could be achieved by training the model on comments from TikTok, which is a younger social media platform. By matching the training population with the target population, the SAI deployment will be more accurate; however, this also creates another cause for the collection of data.

Surveillance tends to typically accompany some amount of pretext or potentially legitimate argument as justification, so one must inspect the use case for the proposed inclusion of vulnerable classes. Specifically, one must consider whether the stated objectives can be reliability achieved through the proposed SAI use—otherwise, there is low mission integrity. When determining who is surveyed consider: (1) Who was the technology designed to survey? (2) Who is incidentally being observed due to association or proximity? and (3) What are the sources of data that are trained or fed to the algorithm in question?

Methods and instrumentalities

The third axis focuses on the exact technology being deployed. AI deployments include natural language processing for analyzing social media threats to seek threats, machine learning to monitor soil samples or meteorological readings to better predict environmental conditions, and computer vision algorithms to determine the identity of criminal suspects in images. It is key to note that many of these deployments may incorporate more than one method of artificial intelligence and human oversight in deployments with various

levels of "human in the loop" (the extent of human supervision or oversight) (Wang, 2019). To get a holistic image of the various levels of technology and human supervision, one must observe the system being operated by a real end user to understand all of the interactions.

Function creep is the relevant aspect here. Technology must not be viewed as a stable and stagnant application, you must assume global developments will soon occur that will alter the initial form of the SAI. Entities with access to a large amount of capital or talent are the most susceptible to function creep in the methods and instrumentalities space. Function creep is not inherently bad; after all, it's what led us from inactivated vaccines to mRNA vaccines, which can address new classes of disease (Dolgin, 2021). It is the blind function creep that morphs existing technology into a new form and deploys it without adequate assessment or notice. Additionally, with respect to the private vendors and government users, this is the aspect least likely to be understood by the surveyor as it's the most technically complex.

A vendor/creator must also elect how much control to give the end user. In facial recognition, some systems permit the law enforcement end user to adjust aspects, like confidence intervals, to more readily influence algorithm output (Farivar, 2018). This frequently results in an untrained field end user being able to see a variety of outputs (e.g., images of suspects) without fully understanding the context and true meaning of the results. Therefore, entire databases of millions of people can be searched against a probe image with the suspect (if they are present in the database) and those that most closely resemble the suspect receiving the most scrutiny (even if statistically they should be excluded or receive less scrutiny). This is most dangerous when forces (e.g., confirmation bias) are at play that turn objective outcomes into supporting biased evidence. The result is that more people are often surveyed than necessary.

When determining the methods and instrumentalities consider: (1) What is the core underlying technology that the Surveillance AI is relying on? 2) How does this specific technology and vendor compare to the general market in terms of performance? and 3) Is there sufficient technical safeguards and human or automated oversight, and what may be needed in the future?

Passive or real time

SAI is typically used on a spectrum of supervision/autonomy that is dictated by the surveyor's objectives and dependency on the system. In one scenario, the output of the system is immediately factored into some decision making or action. The benefit of real-time systems is that certain processes can be automated. Most frequently, these occur in instances of law enforcement or public safety. An agency responsible for implementing the SAI, such as ICE at the United States border, may use an automated drone to detect undocumented persons crossing the border and send human officers to detain them (Mascellino, 2021). Real-time systems are also commonly found in smart cities where the environment is supported by a high volume of sensors and cameras, and they are connected through a cloud-like storage system. One proposed and politically neutral SAI application is reducing traffic congestion by implementing the technology into traffic lights. Real-time data is also useful for medical diagnosis or surveillance where situations are rapidly evolving and fluctuating variables are present, like AI supported contact tracing (Hutchins, 2020).

The vulnerability of real-time systems is that they rely on a set of assumptions. These include that all the environment's key factors are accounted for (there are no outliers); that the predictive factors determined to be successful will maintain their accuracy over time; and that the user is interpreting the readings in a predictable, accurate, and consistent manner (sufficient training). Even if these three concerns can be addressed, real-time systems are extremely susceptible to mission creep because they are often built to be adaptable and applied to new situations or persons.

If the information is passive then typically a human user is initiating the operation or process on each instance. An example would be a social benefits agency investigating someone for potential fraud. While there is an existing model the operator may have to upload the person's specific financial, social, or economic profile for consideration. Additionally, they may use other sources of information, like the individual's testimony or a private investigator, instead of relying on the information in a vacuum. While the ability of an existing passive system to transform into real time is only possible through function creep, mission creep is also the relevant issue here because it often requires redefining of the project's goals and capabilities. Given the sophistication of technology vendors, they often offer tiers of service with the high-end tiers including real time services and the low-end tiers including passive, less sophisticated systems.

When assessing passive or real time consider: (1) How much is the information being relied on? (2) Does the system automatically process new information or does that need to be initiated? and (3) What safeguards exist should bad data be entered into the system?

Subsequent effect

The subsequent effect relates to the direct results of using the surveillance system. Questions of subsequent effect are more traditionally determined by the organization deploying the technology and less so by the original technologists. Some higher stakes governmental use cases include a police officer trying to identify an unidentified person, a public health official who is notified of someone's exposure to COVID-19 or a positive COVID-19 test result, and a federal agent who uses SAI to gain evidence required for a warrant. Subsequent effect is what that agent or organization *does* with that information. Systems where the output is corroborated with other evidence or information before a decision is made or an action is taken are much more sound and robust as they recognize the ability for the technology to falter and produce false results. In determining adequate safeguards, one must have an understanding of the performance metrics of the system to determine how much to rely on the results. Common metrics of SAI performance include confidence intervals, false-positive rates, and t-values. While these values can determine technical soundness, it is important to consult with domain experts for the applied field such as public health officials in the contact tracing app to ensure there are no unintended subsequent effects or consequences.

The surveyor is the key party in determining the subsequent impacts. When determining the subsequent effect consider: (1) What was the original purpose of the technologies implementation and what is the duty of the entity operating it? (2) Who has access to the outputs and could they ever use this information in an unintended manner? and (3) What are the unintended consequences of using this system?

Accountability

Accountability refers to the integrity of the management of the system. It is a measure of faith that the surveyed have in the surveyors that the tools deployed are not being used arbitrarily, maliciously, negligently, or otherwise below a standard of professionalism and ethics. Occurrences of these factors not only have the capability of undermining faith in that specific SAI system but also in the institution as a whole. American law enforcement has a history of misusing certain databases or tools for purposes outside of the original intended use. In California, over 1,000 cases of computer database misuse by law enforcement agencies have been identified in the last decade (Stanton et al., 2019). The reasons include stalking women and running covert background checks. These instances are rarely fully prosecuted.

When evaluating accountability, consider: (1) Who is responsible for consistently ensuring the results are not inaccurate, biased, or inequitable and what resources are supporting this effort? 2) Do the creators and vendors of this proposed system have a good track record of honesty and moral character? 3) What safeguards exist within an organization to ensure the tool is only used for its intended purpose (e.g., auditing functionality and restriction of access)?

Transparency

Transparency describes the relationship between the surveyor and the surveyed. It's a measurement of the effort and intentionality the surveyor has put into expressing the existence, functioning mechanisms, implications, and persons associated with an SAI system. It is not merely the surveyed people being aware it exists; many SAI systems are discovered by journalists or external parties. The information is not being willfully disclosed in this scenario. There are situations however of national security where the covert deployment of SAI may be required to maintain the element of surprise. In non-edge cases, however, it is generally permissible to share the type of SAI being deployed, the vendor being used, the type of financing/use arrangement, the skill or education of the operators, and the overall goals of deploying the system. Ideally, this information should be provided before deployment occurs, during the pitch or pilot stage.

Questions to consider when evaluating transparency include: (1) Are the individuals subjected to the SAI aware of its presence and impact? 2) Do key decision makers who rely on outputs from the system have enough information to contextualize outputs? and 3) Have any conflicts of interest been addressed and disclosed to the surveyed and any key stakeholders?

Trends in Surveillance AI

A summary of several pervasive trends in Surveillance AI can be found in Figure 39.2 below.

Trend	Parties Involved	Definition	Example
Disparate Deployment and Impact	Surveyors, users, and creators of SAI	Analog and modern surveillance have always focused on marginalized groups and are less likely to be deployed in more affluent areas. Inner cities tend to have a higher density of cameras and surveillance tools deployed.	Israeli government is using a facia l recognition program called Blue Wolf to survey Palestinians and determine whether to arrest, detain, or leave them undisturbed(Roth 2021).
Cybersecurity Vulnerability	Surveyors and creators	Deployed systems face two critical issues, being attractive targets (due to possession of highly structured and labeled data) and low cybersecurity preparedness.	The Customs and Border Protection agency (CBP) was running a facial recognition program on travelers and 100,000 images of foreign travelers were stolen (Harwell and Fowler 2019).
Predatory Procurement Practices	Vendors	Vendors in this space have frequently misrepresented the efficacy and origins of their technology, as well as engaged in illegal or ethically questionable behavior.	Banjo, the recipient of a $20 million surveillance system built in Utah, failed to deliver on real-time surveillance capabilities and instead was indicated to have just the capability to aid a skilled analyst through a dashboard interface (Office of the St ate Auditor 2021). The auditors specifically stated that Banjo failed on core claims, did not

FIGURE 39.2 Summary of SAI trends, definitions, examples, and parties involved

			include artificial intelligence, and that those state officials would have detected the deficiencies, had they not relied merely on the claims of the technology (Rodgers 2021).
Privacy Advocacy	Government, Policymakers, and Academics	Differing from the previous era's privacy legislation like HIPPAA and FERPA, this legislation focuses less on the identity of the individual (health care recipient and student respectively) and more so focuses on the type of information sought to be regulated.	The Biometric Information Protection Act of Illinois focuses on protecting the rights of Illinois residents as it pertains to the collection of biometric information and places emphasis on consent and notification of data collection processes. Facebook violated BIPA by defaulting users into facial recognition and recently settled in for $650 million (Bryant 2021).
Countersurveillance and Crowdsourced Resistance	The Surveyed, private technologists, and nonprofits	To combat the rapid expansion of SAI, individuals and groups have been building technical tools that disrupt SAI, track and publicize otherwise covert systems, and employing a variety of other means.	Amnesty International's Decode Surveillance NYC have been supported by thousands of volunteers to track New York City's use of facial recognition (Amnesty International 2021a).Participants simply mark the physical presence of official camera NYC cameras on a map as they go about their daily life

FIGURE 39.2 Continued

Belt and Road Initiative	China and Developing Nations	One of the biggest criticisms of China's Belt and Road Initiative is that it's been directed at developing nations that do not currently have reliable infrastructure. A common tactic is to offer the infrastructure or technology as a debt instrument that takes years to pay back, or allow the foreign institution to operate as a market participant in the utility. These vendor companies will permit China's influence to be exerted through a critical international infrastructure need for years to come.	China has specifically built power plants and grid infrastructure in over 65 countries throughout Asia, Southern Europe, and Eastern Africa (Feng et al. 2020). Nations like Thailand, Sudan, and Venezuela have all made significant relationships with the Belt and Road Initiative and

FIGURE 39.2 Continued

Mitigating Surveillance AI Harms and Creating a Privacy Preserving Society

There are a number of law- and policy-based measures governments can take to address the harms presented by SAI.

Efficacy and security

Efficacy and security concerns center on whether SAI meets advertised and discussed parameters, as well as whether a system is designed and operated in a way that minimizes the risk of unauthorized access. In short, efficacy and security concerns revolve around whether a given SAI system can be responsibly used in a given deployment environment. Systems are high priority targets for bad actors and often lack sufficient cybersecurity measures to limit the amount and severity of breaches. SAI hacking can lead to theft of personally identifiable information, other privacy breaches, and ransomware attacks. Given the deficit of technical expertise in most government agencies, efficacy and security concerns are usually best addressed by technology vendors and creators—although a public mandate may be required for them to do so.

One solution would be to empower and fund the Federal Trade Commission (or its international equivalent) with resources to study and produce rules regarding SAI marketing and sales. These funds would enable the FTC to study SAI, catalog best practices, and conduct rulemaking governing use. The root of the problem stems from over ambitious and, at times, fraudulent claims of the performance of the technology. Taking these actions will better protect governmental consumers of SAI and ensure that purchased technology lives up to the stated performance. Specifically, the Office of Technology Research and Investigation within the FTC should take primary responsibility for this directive, given their combination of technical expertise and mandate for consumer protection. The Office's efforts should be carried out in consultation with the National Institute of Science and Technology (NIST, a non-regulatory agency that has previously studied many types of SAI) and the National Artificial Intelligence Initiative (NAII).

Transparency and ethics

State and local deployments of SAI tend to be more problematic given the lack of technical expertise frequently found in the Federal government. Given this trend, a Privacy Pilot Program would allow state and local municipalities to use a tailored hybrid model of government and civilian oversight for Surveillance AI. The progression of this technology in the United States has demonstrated that municipalities and jurisdictions may be better served using proactive rather than reactive methods to address SAI use. The goal of the program would be to develop effective and tailored strategies for deploying SAI in ways that respect the civil rights and privacy of city residents, and to implement frameworks that assure responsible continued use of the technology moving forward. Members of the pilot program should receive economic support and resources to establish one or a combination of the following:

- Privacy Commission. A board of majority internal city employees who have veto, rulemaking, and investigative powers over the use of Surveillance Technology by any government office or employee in their jurisdiction. The Privacy Commission would also draft surveillance ordinances, when appropriate, for consideration by the City Council or similar body.
- Digital Privacy Task Force. A board comprising internal employees, academics, industry leaders, advocacy groups, stakeholders, and community members who would serve in an advisory capacity, helping the city craft broad privacy policy goals as well as respond to specific use cases or applications.
- Citizen Oversight Board. A board of community members and leaders who regularly confer with law enforcement over the deployment of SAI and novel surveillance tools.
- Privacy Office. A government office, overseen by a Chief Privacy Officer (CPO) and staffed by analysts, policy advisors, and program managers, tasked with reviewing city policy around privacy and the use of SAI. Cities that currently have a CPO or existing privacy office could use program funds to expand their staff, purchase training tools and resources, and deploy privacy-preserving software or methods.

A Brave New World

SAI use has immediate and long-term impacts on all communities that come in contact with the technology. While the end goal may be a justifiable or otherwise positive application, any myriad of technical, logistical, or sociological impacts can frustrate the purpose of the deployment. There is currently a severe lack of oversight and safeguards to ensure the ethical deployment of the technology. Responsibility for monitoring SAI deployments must involve a wider group of stakeholders. By using the two concepts of mission creep and function creep, the Seven Axes of Surveillance Framework, observing the common trends of SAI, and integrating democratic oversight mechanisms, we can more readily address problematic SAI use and even prevent future use cases from being deployed. It is only by taking a multifaceted approach that the harms can be mitigated and any benefits from SAI can be fully realized.

Further complicating the notion of government use of SAI are three ever present factors. The first is vendor reliance. Due to the arduous and time intensive process surrounding government procurement contracts, municipalities are often locked into contracts with the wrong vendor but left with little recourse or resources to make an alternate selection. The second is the pervasive issue of privacy, which is also at play when these large troves of information are being assembled. Finally, the issue of opportunity cost and the potential for a more efficient allocation of resources is always at play. The efficacy of many forms of SAI and their impact on crime rates are hotly debated due to large disagreement regarding what elements actually contribute to crimes.

The ever-expanding SAI state creates a system where one must ask not only if the use is justified, but also if the creation of the *technology* is equally justifiable. While a Pandora's box of data collection processes has been opened and these practices are permissible by many governments, there remains a vital ethical question. A similar question is at the heart of the debate over the raw materials for consumer electronics as well as electronic batteries. Can the creation of an SAI system that is dependent on a collection featuring millions of instances with a lack or no informed consent be truly justified?

References

American Civil Liberties Union. (n.d.). Background on the State Secrets Privilege. American Civil Liberties Union. https://www.aclu.org/other/background-state-secrets-privilege.

American Civil Liberties Union. (2021). Surveillance under the USA/PATRIOT Act. American Civil Liberties Union. https://www.aclu.org/other/surveillance-under-usapatriot-act.

Amnesty International. (2021a). Digital volunteers to expose scale of facial recognition technology in New York City. Amnesty International. May 4. https://www.amnesty.org/en/latest/news/2021/05/decode-surveillance-nyc-launches/.

Amnesty International. (2021b). New York is in danger of becoming a total surveillance city. Amnesty International. June 3. https://www.amnesty.org/en/latest/news/2021/06/scale-new-york-police-facial-recognition-revealed/.

Bowles, Nellie. (2020). Why is a tech executive installing security cameras around San Francisco? *The New York Times*, July 13. https://www.nytimes.com/2020/07/10/business/camera-surveillance-san-francisco.html.

Brown, Dalvin. (2020). The U.S. government is "looking at" banning TikTok due to national security concerns. *USA Today*, July 7, pp. 1–2. https://www.usatoday.com/story/tech/2020/07/07/u-s-looking-at-banning-tiktok-cites-chinese-surveillance/5390430002/.

Brown, Madeline, & Ezike, Richard. (2020). How governments, nonprofits, and the private sector can help technology work for everyone. Urban Institute. June 26. https://www.urban.org/urban-wire/how-governments-nonprofits-and-private-sector-can-help-technology-work-everyone.

Bryant, Jennifer. (2021). Facebook's $650M BIPA settlement "a make-or-break moment". International Association of Privacy Professionals. March 5. https://iapp.org/news/a/facebooks-650m-bipa-settlement-a-make-or-break-moment/.

Campbell, Ian Carlos. (2021). UK police surprised to learn energy-intensive weed farm is actually a bitcoin mine. *The Verge*, May 28, pp. 1–2. https://www.theverge.com/2021/5/28/22458632/bitcoin-mine-mistaken-cannabis-growing-operation-police.

Chao, Becky, Null, Eric, Collins-Dexter, Brandi, & Park, Claire. (2019). Centering civil rights in the privacy debate. *New America*, September 17, pp. 1–13. https://www.newamerica.org/oti/reports/centering-civil-rights-privacy-debate/for-marginalized-communities-the-stakes-are-high/.

Cox, Joseph. (2020). How the U.S. military buys location data from ordinary apps. *Vice.com*, November 16. https://www.vice.com/en/article/jgqm5x/us-military-location-data-xmode-locate-x.

Crump, Catherine. (2016). Surveillance policy making by procurement. *Washington Law Review 91*(4), 1595. https://digitalcommons.law.uw.edu/wlr/vol91/iss4/17/.

Dai, Sarah. (2019). China's face-scan tech now stretches to trash cans and public housing. *South China Morning Post*, August 2, pp. 1–2. https://www.scmp.com/tech/policy/article/3020977/chinas-facial-recognition-mania-now-extends-public-housing-and-trash.

Dolgin, Elie. (2021). The tangled history of MRNA vaccines. *Nature 597*(7876), 318–324. https://doi.org/10.1038/d41586-021-02483-w.

Durbin, Kaitlin. (2019). Utility companies, neighbors help police find illegal Colorado marijuana grows. *Colorado Springs Gazette*, August 5, pp. 1–4. https://gazette.com/crime/utility-companies-neighbors-help-police-find-illegal-colorado-marijuana-grows/article_f5bfe0a5-308e-5980-b0c9-5e664ddfc005.html.

Farivar, Cyrus. (2018). Amazon: Cops should set confidence level on facial recognition to 99%. *Ars Technica*, July 30, pp. 1–3. https://arstechnica.com/tech-policy/2018/07/amazon-cops-should-set-confidence-level-on-facial-recognition-to-99/.

Fassler, Ella. (2021). Oklahoma quietly launched a mass surveillance program to track uninsured drivers. *OneZero*, April 7, pp. 1–5. https://onezero.medium.com/oklahoma-quietly-launched-a-mass-surveillance-program-to-track-uninsured-drivers-471bb4e5701a.

Feng, Tian-tian, Gong, Xiao-lei, Guo, Yu-hua, Yang, Yi-sheng, Pan, Bin-bin, Li, Shao-ping, & Dong, Jun. (2020). Electricity cooperation strategy between China and ASEAN countries under "the Belt and Road". *Energy Strategy Reviews 30*(July), 100512. https://doi.org/10.1016/j.esr.2020.100512.

Bedoya, A., Frankle, J., & Garvie, C. (2016). *The perpetual lineup: Unregulated police face recognition in America*, pp. 1–11. https://www.perpetuallineup.org/.

Gellman, Barton, & Adler-Bell, Sam. (2017). The disparate impact of surveillance. *The Century Foundation*, December 21. https://tcf.org/content/report/disparate-impact-surveillance/.

Georgetown Law. (n.d.) The perpetual lineup: Unregulated police face recognition in America. https://www.law.georgetown.edu/privacy-technology-center/publications/the-perpetual-line-up/.

Georgetown Law. (2019). The color of surveillance: Monitoring of poor and working people. https://www.law.georgetown.edu/privacy-technology-center/events/color-of-surveillance-2019/.

Greene, Robyn. (2018). How cities are reining in out-of-control policing tech. *Slate Magazine*, May 14. https://slate.com/technology/2018/05/oakland-california-and-other-cities-are-reining-in-out-of-control-police-technologies.html.

Harwell, Drew. (2019). FBI, ICE find state driver's license photos are a gold mine for facial-recognition searches. *Washington Post*, July 7. https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/.

Harwell, Drew, & Fowler, Geoffrey. (2019). U.S. Customs and Border Protection says photos of travelers were taken in a data breach. *The Washington Post*, June 10. https://www.washingtonpost.com/technology/2019/06/10/us-customs-border-protection-says-photos-travelers-into-out-country-were-recently-taken-data-breach/.

Holt, Kris. (2021). NYPD secretly spent $159 million on surveillance tech. *Engadget*, September 10, pp. 1–4. https://www.engadget.com/nypd-surveillance-slush-fund-special-expenses-facial-recognition-stingray-155039463.html.

Hutchins, Todd E. (2020). The legality of artificial intelligence contact tracing to stop coronavirus in the U.S. *Yale Journal of Law and Technology*, September 12, pp. 1–7. https://yjolt.org/blog/legality-artificial-intelligence-contact-tracing-stop-coronavirus-us.

International Dark-Sky Association. (2021). Black history month: A history of forced illumination. International Dark-Sky Association. February 18. https://www.darksky.org/black-history-month/.

Keck, Catie. (2019). Say cheese! EU votes to amass a giant database of biometric data. *Gizmodo*, April 22. https://gizmodo.com/say-cheese-the-eu-is-set-to-create-one-of-earths-bigge-1834225890.

Khan, Saher, & Ramachandran, Vignesh. (2021). Post-9/11 surveillance has left a generation of Muslim Americans in a shadow of distrust and fear. *PBS NewsHour*, September 16, pp. 1–6. https://www.pbs.org/newshour/nation/post-9-11-surveillance-has-left-a-generation-of-muslim-americans-in-a-shadow-of-distrust-and-fear.

Knowledge Sourcing. (2021). Global data broker market size & share: Industry report, 2021–2026. Knowledge Sourcing Intelligence LLP. May, p. 1. https://www.knowledge-sourcing.com/report/global-data-broker-market.

Lyons, Kim. (2021). Amazon's Ring now reportedly partners with more than 2,000 US police and fire departments. *The Verge*, January 31, pp. 1–2. https://www.theverge.com/2021/1/31/22258856/amazon-ring-partners-police-fire-security-privacy-cameras.

Manjoo, Farhad. (2012). The outrageous conservative claim that the government didn't invent the internet. *Slate Magazine*, July 24. https://slate.com/technology/2012/07/who-invented-the-internet-the-outrageous-conservative-claim-that-every-tech-innovation-came-from-private-enterprise.html.

Mansoor, Sanya. (2021). Muslim Americans bring the fight against surveillance to the Supreme Court. *Time*, September 16, pp. 1–10. https://time.com/6097712/muslim-american-surveillance-supreme-court-sept-11/.

Mascellino, Alessandro. (2021). ICE drops $3.9M for trust stamp's facial recognition, London Met taps NEC for $4.2M. *Biometric Update*, September 27, pp. 1–3. https://www.

biometricupdate.com/202109/ice-drops-3-9m-for-trust-stamps-facial-recognition-london-met-taps-nec-for-4-2m.

National Immigration Law Center. (2016). How ICE and DMVs share information. National Immigration Law Center. May 2016. https://www.nilc.org/issues/drivers-licenses/ice-dmvs-share-information/.

News from Elsewhere. (2019). Chinese driver gets ticket for scratching his face. *BBC News*, May 24. https://www.bbc.com/news/blogs-news-from-elsewhere-48401901.

Nouri, Steve. (2020). Council post: How AI is making an impact on the surveillance world. *Forbes*, December 4, pp. 1–7. https://www.forbes.com/sites/forbestechcouncil/2020/12/04/how-ai-is-making-an-impact-on-the-surveillance-world/?sh=3cbec00e265e.

Office of the State Auditor. (2021). Review of Banjo IT Controls memo v14. https://reporting.auditor.utah.gov/servlet/servlet.FileDownload?file=0151K000004219lQAA.

Owen, Richard. (2019). Are universities fuelling surveillance capitalism? *Digital Societies*, November 4, pp. 1–7. https://www.digitalsocieties.co.uk/blog/are-universities-fuelling-surveillance-capitalism.

Pivcevic, Katya. (2020). EU border agency says biometrics interoperability framework includes robust safeguards. *Biometric Update*, December 29, pp. 1–6. https://www.biometricupdate.com/202012/eu-border-agency-says-biometrics-interoperability-framework-includes-robust-safeguards.

Privacy International. (n.d.). Visual surveillance technology. Privacy International, p. 1. https://privacyinternational.org/learn/visual-surveillance-technology.

RegLab. (n.d.). Tax enforcement—RegLab. The Regulation, Evaluation, and Governance Lab. https://reglab.stanford.edu/projects/tax-enforcement/.

Robertson, Adi. (2014). Oakland votes to limit scope of centralized surveillance program. *The Verge*, March 5, pp. 1–3. https://www.theverge.com/2014/3/5/5473722/oakland-votes-to-limit-scope-of-domain-awareness-center.

Rodgers, Bethany. (2021). State auditors find that Banjo oversold its surveillance tech to the Utah Attorney General's Office. *The Salt Lake Tribune*, March 30, pp. 1–3. https://www.sltrib.com/news/politics/2021/03/30/state-auditors-find-that/.

Roth, Emma. (2021). The Israeli army is using facial recognition to track Palestinians, former soldiers reveal. *The Verge*, November 8, pp. 1–2. https://www.theverge.com/2021/11/8/22769933/israeli-army-facial-recognition-palestinians-track.

Sato, Mia. (2021). Singapore's police now have access to contact tracing data. *MIT Technology Review*, January 5, pp. 1–3. https://www.technologyreview.com/2021/01/05/1015734/singapore-contact-tracing-police-data-covid/.

Schiffer, Zoe. (2020). Google fires prominent AI ethicist Timnit Gebru. *The Verge*, December 3, pp. 1–4. https://www.theverge.com/2020/12/3/22150355/google-fires-timnit-gebru-facial-recognition-ai-ethicist.

SOMA West. (n.d.). What is a community benefit district?—The SOMA West Community benefit district. SOMA West CBD, p. 1. https://somawestcbd.org/what-is-a-community-benefit-district/.

Stanton, Sam, Smith, Darrell, & Wailoo, Elliot. (2019). California's criminal cops: Hundreds of officers misuse law enforcement databases. *The Mercury News*, November 12, pp. 1–6. https://www.mercurynews.com/2019/11/12/californias-criminal-cops-hundreds-of-officers-misuse-law-enforcement-databases/.

Steinberger, Michael. (2020). Does Palantir see too much? *New York Times*, October 21, pp. 1–27. https://www.nytimes.com/interactive/2020/10/21/magazine/palantir-alex-karp.html.

Wang, Ge. (2019). Humans in the loop: The design of interactive AI systems. Stanford Institute for Human-Centered Artificial Intelligence. October 21. https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

Wheeler, Brian. (2016). Police surveillance: The US city that beat big brother. *BBC News*, September 29. https://www.bbc.com/news/magazine-37411250.

Whittaker, Zack. (2020). Data brokers track everywhere you go, but their days may be numbered. *TechCrunch*, July 9, pp. 1–3. https://techcrunch.com/2020/07/09/data-brokers-tracking.

Zubair, Ayyan. (2019). Automated license plate readers & law enforcement. S.T.O.P.—the Surveillance Technology Oversight Project. July 5, pp. 1–5. https://www.stopspying.org/latest-news/2019/7/5/automated-license-plate-readers-amp-law-enforcement.

CHAPTER 40

SMART CITY TECHNOLOGIES

A Political Economy Introduction to Their Governance Challenges

BEATRIZ BOTERO ARCILA

Introduction

"Smart city technologies" is a term used to refer to the concerted move over the past two decades to network urban infrastructures and to use computation to solve urban problems and deliver city services more efficiently.

The term refers both to computational models of urbanism and to data-driven and algorithmically intermediated technologies principally catered to municipal governments and entities providing city services (Mattern, 2021; Kitchin, 2014). So-called smart city technologies typically rely on a base of smartphones and other sensors connected by high-speed communication networks and applications or interfaces that translate data into alerts, insights, or information for government operators or final users (McKinsey Global Institute et al., 2018). Smart city technologies are designed to deliver new efficiencies, insights, and conveniences for both citizens and city-governments alike: these tools are as varied as waste- and utility-management systems, public transportation apps, smart streetlights, bike-sharing systems, security cameras, centers of command and control, and early alert systems.

Although there are still no—or if so, very few—cities that have built complete digital twins or that are wired at all levels of their infrastructure, smart city technologies are becoming ubiquitous and common. They are often presented by their corporate and government champions as having a critical role in facilitating innovation in local government, improving transportation and infrastructure, and preparing cities to respond to environmental and energy challenges, among others. They also promise to "enable economic, social and environmental sustainability and transform communities" and to create a platform for innovation and entrepreneurship (Baykurt & Raetzsch, 2020; US Department of Transportation, 2017; Rao, 2010; McKinsey Global Institute et al., 2018).

On the other side, information scholars have pointed out that smart city technologies, like many other digital technologies, can entrench inequality and discrimination patterns

in cities' fabric and turn cities into surveillance complexes. These risks arise especially in instances in which these tools are involved in decision-making processes that don't have right or wrong mathematical answers but involve inherently political and policy questions (CIHR, 2015). This is the case, for example, of predictive policing tools that tend to flag minorities as likelier of being involved in criminal activities, and less so of public utilities optimization tools.

This chapter is an introduction to the governance challenges smart city technologies pose. It focuses on the inequality, discrimination, and surveillance challenges. As such, the chapter does not spend a lot of time emphasizing the benefits and potential of these technologies—for example, that they can help cities to adapt to the climate crisis or that they can improve urban policymaking and the delivery of city services (for more of that see Williams, 2020). The chapter takes a political economy approach to technology, which emphasizes how these tools, and their distributive effects, are often the product of the logic of the market, law, politics, and ideology. By doing so, this chapter argues, the challenges posed by smart city technologies should be understood as deeply intertwined with the incentives that different actors face during the production and adoption cycles of these technologies. Similarly, the realization of their benefits will depend on the shape of the institutions, actors, and infrastructures that sustain them. This should help interested parties add some nuance to binary analyses and identify different actors, institutions, and infrastructures that can be instances of intervention to shape their effects and create change.

This chapter proceeds as follows: It begins by offering a brief overview of the literature on smart city technologies and two opposing narratives—one focusing on their efficiency-related benefits, and another focusing on the risks they pose. The chapter then dives deeper into the digital technologies that are adopted in cities to provide security services, largely in the context of policing, and highlights how ideology, the market, and institutions—the elements of a political economy framework—influence the design, adoption, and effects of these technologies.

Before we continue, a small clarification is due. The terms "smart cities" or "smart city technologies" are not uncontroversial. Smart cities is a term that was first put forward by corporate players to push for the vast adoption of these technologies in the late 2000s. Many have pointed out that the term presents cities as something in which a particular form of efficiency should be a core value, emphasizing and embodying the interests and logics of corporate service providers and obscuring how complex cities are as social, spatial, economic, and legal phenomena (Green, 2019; Mattern, 2021). At the same time, it seems to ignore that despite the relative novelty of smart city technologies, and the vast scale of the information that can be amassed today, city planners and city officials have been using data analytics and forecasting tools for about a century (Brayne, 2020; Williams 2020). The efficiency-based critique and the term itself also often ignore that efficiency has been at the core of urban policymaking and planning, municipal law, and the economic theory around cities since at least the mid-twentieth century in most of the western world (Botero Arcila 2021; Frug 1980). Thus, like all definitions, the terms "smart cities" and "smart city technologies" are doing a certain form of work, one that advantages a particular relationship between city government and corporate power, and obscures other histories and legacies. As such, it is setting a frame for how these tools will be understood, measured, valued, and governed.

While being aware of the term's politics and limitations, I choose to use it in this chapter mainly because it continues to be the main way to refer to these industrial formations and

the technologies particular and common to the urban governance context. The term, as I use it here, includes the algorithmic breakthroughs, the product improvements, and the greater efficiency and convenience, while throughout the chapter I emphasize that they are also technical and social practices, shaped and sustained by institutions and infrastructures.

Two Tales of Smart Cities

The corporate narrative about the opportunities of smart city technologies sets the terms for much of the debate and conversation about smart city technologies. The term comes from IBM's 2008 "Smart Cities" and "Smarter Planet" advertising campaigns to promote the use of technology and data to analyze problems of cities (IBM, n/a). Not too long afterwards, in 2010, Cisco launched its now closed "Smart Connected Communities" program, which was based on using data analysis and web-based interface programs to connect cities through technology (Cisco, n/a). This section presents, first, this corporate narrative, and then some of the main critiques and warnings raised by information scholars about this narrative and the consequent adoption and deployment of these tools.

The corporate narrative and patterns of adoption

According to Cisco, "a smart city uses digital technology to connect, protect, and enhance the lives of citizens. IoT sensors, video cameras, social media, and other inputs act as nervous system, providing the city operator and citizens with constant feedback so they can make informed decisions" (Cisco, n/a). A report by the McKinsey Global Institute similarly explains that a smart city is a city that has a technology base that includes a critical mass of smartphones and sensors connected by high-speed internet and open data portals. It includes specific applications that translate that data into alerts and city-related insights that suggest certain actions from the application users and from the local government (McKinsey Global Institute et al., 2018). The point is that the vast troves of data these technologies collect can provide surprising insights about cities, and can revolutionize how city services are planned and delivered with increasing efficiencies (Ash Center for Democratic Governance and Innovation, 2017; Goldsmith & Crawford, 2014).

Indeed, although cities have been using data for a while, the sheer and vastness of the data that can be collected and analyzed today is unprecedented. At the same time, these technologies have become easy for cities to adopt and access and are sufficiently reliable when deployed with care (Williams, 2020). Consequently, most large cities in the world have adopted some form of digital technology to plan and deliver their city services. The McKinsey report (2018) found, for example, that the cities that have the largest extents of sensors and devices, the best communication networks, and open data portals are large world cities like Amsterdam, New York, Seoul, Singapore, Stockholm, Shenzhen, and Singapore. Although cities in the global south lag in this ranking, large cities like Sao Paulo, Buenos Aires, Abu Dhabi, Dubai, and Bangkok had also implemented smart city applications to some extent. Mobility and security services are some of the city-realms

where these technologies have been most widely adopted (McKinsey Global Institute et al., 2018).

In the local policing context, for example, big data policing tools collect, integrate, analyze, and share data sources that law enforcement has access to and then narrow down searches for suspects or potential sites with higher risks of crime and alert officers where to intervene. "Heat lists" are an instrument increasingly used by large cities' police departments to identify individuals who are at higher risk of being a victim or perpetrator of violence (Rhee, 2016). Some lists draw from elements like criminal history, arrests, or whether a person has been identified as part of a gang to identify youth that are likely to be involved in a violent act—as a victim or a perpetrator (Chicago Police Department, 2015). Other examples of heat lists, as one being developed in Medellin, Colombia, seek to identify teens at high risk of sexual abuse to prevent child and teenage pregnancy. Using them, police departments can better plan how to deploy their forces or public officials can send a social security worker to a girl at risk before it's too late (Atehortua, 2021).

The report by McKinsey finds, however, that even the front-runners lacked a comprehensive technology infrastructure, and had varied degrees of awareness, use, and satisfaction (stronger in Asia, and less so in European cities, for example) (McKinsey Global Institute et al., 2018) Similarly, a 2020 study found that smart cities are not yet the responsive and hyper-efficient city that most technology companies present. In most cases, residents don't engage much with these technologies and, in practice, the impact of data on the delivery of urban services is still relatively limited (Gufflet & Kemp, 2021).

The critiques to and the risks of the corporate narrative

Several critiques have been raised in response to the baseline story that suggests that smart city technologies can neutrally improve city services. Critics point out the fact that these technologies can aggravate, or at least reproduce, racial discrimination and socio-economic inequality, and that they can entrench surveillance in our cities' fabric, eroding privacy and other related civil liberties. Another common critique is that the allure of these tools, often designed with corporate interests in mind, can get prioritized over better, integral solutions to local issues. In the contexts of policing, for example, the critics of these technologies show that the adoption of these tools can incentivize police departments to ignore the complexities of local crime, prioritize control over prevention, enhance the surveillance and policing of already marginalized communities, and make local governments ever more dependent from technology companies (Green, 2019; Brayne, 2020). In what follows, I will describe these three main risks: (1) entrenching discrimination and inequality patterns; (2) enhancing surveillance and the subsequent erosion of civil liberties; (3) and the overreliance on technology to solve complex issues.

The risk of entrenching inequality and discrimination patterns

Many of the potentially harmful effects of smart city technologies are related to wider issues posed by data and algorithmic decision making. Indeed, an underlying assumption of the corporate narrative, especially during its earlier days, was that big data would largely

transform how we understand the world: the ideal policy- and decision-making processes would go from testing hypotheses and understanding cause and effects to a world where massive amounts of data and applied mathematics would allow data scientists to draw enough correlations to know what things happen, without really knowing why (Anderson, 2008). In the city context, decision makers would only need to have enough data about cities' behavior to make better decisions about cities, regardless of causes, effects, and context. However, information scholars have, for almost as long, shown that claims to objectivity and accuracy are misleading, especially when these decisions involve humans. Further, the assumption that data is, per se, objective can lead to a variety of unjustified harms (Boyd & Crawford, 2011). Missing out on context, history, and theory often leads to situations in which the adoption of technologies leads to decisions that replicate, without necessarily being aware of it, existing biases and discrimination patterns.

In the US, for example, it has been established that the use of decision-support software in judicial decisions "uncovered evidence of racial bias, finding that when the effect of race, age and gender was isolated from criminal recidivism risk, 'black defendants were 77% more likely to be pegged as at higher risk of committing a future crime and 45% more likely to commit a future crime of any kind" (US Executive Office of the President, 2016). The same is true in other fields. In the banking sector, low-income students did not receive student loans for being considered too risky based on the place they lived, which limited their access to education and opportunity (O'Neill, 2016). Part of these outcomes are based on historic data: because a substantial part of the human-made decisions that were input to the data sets might have been biased and prejudiced, and consequently inaccurate in the best scenario, the outcomes thrown by the algorithms were inaccurate in the same way, even if the mathematical analysis of the data set is sound (O'Neill, 2016).

The same type of problem appears in the city context. Take as an example, Chicago's heat list. The heat list is an instrument created by the city's police department to identify individuals who are at a higher risk of being a victim or perpetrator of gun violence and it creates risk scores from 1 to 500 (Rhee, 2016). The list draws from elements like criminal history, arrests, or whether a person has been identified as part of a gang (Chicago Police Department, 2015). Using the list, police seek to intervene in an individual's life before the violent act happens. Individuals with high-risk score may receive a letter specifically directed to them, or a senior police officer, a social worker, or a member of the community will visit the individual's home to try to deter them from engaging in violence. The list is rather accurate: A 2016 report showed that 70 percent of people who had been shot in Chicago were on the list, as were more than 80 percent of those arrested in connection with the shootings (Davey, 2016).

There are, however, important concerns about how these predictive policing tools run the risk of becoming self-fulfilling predictions that replicate existing racial biases, which in turn would lead to over-policing communities of color. Because the inputs to these lists include socio-economic factors that correlate with structural racism and poverty, risk-based policing leads to enhanced policing in communities more affected by racism and poverty. This, in turn, leads to more arrests in these communities and even higher risk-scores. The tools thus contribute to a self-reinforcing cycle: because a high-risk score makes it more likely that a judge will hold the defendant in jail without bond, individuals belonging to communities of color or over-policed neighborhoods may end up spending more time in

the carceral system. On the contrary, a low score makes it more likely that the defendant will be released (Benjamin, 2019).

Take as a second example the case of facial recognition technologies. Facial recognition technology is an increasingly popular tool used by law enforcement departments to identify suspects of different crimes. As an AI-powered technology, facial recognition software typically analyzes countless photos of people's faces and learns to make predictions about which images are of the same person. The facial recognition technology being deployed in sensitive contexts, however, is still too flawed to recognize minorities. In 2019, the National Institute of Standards and Technology (Grother et al., 2021) found that many facial recognition algorithms were less accurate in identifying people of color, which means they worsen systemic bias in the criminal-justice system (Porter, 2019). Still, a 2021 report focused on face recognition algorithms applied to confirm airline passenger identities found that several of the algorithms could perform the task using a single scan of a passenger's face with 99.5 percent accuracy or better (Grother et al., 2021). In 2020, nevertheless, there were three cases documented of individuals who were wrongly arrested and briefly jailed based on a bad facial-recognition match in the United States. In all three cases, the individuals were Black men (Hill, 2021).

Not in small part as a response to the risks of discrimination of the technology, at least two dozen cities and states in the United States have banned government use of facial recognition technology (Simonite, 2021). Similarly, at the time of writing, the AI Act being discussed in the European Union contemplates a series of restrictions and safeguards to the use of facial recognition for "real time" remote identification in publicly accessible spaces for the context of law enforcement (European Commission, 2021). Despite this progress, the technology is increasingly adopted in other contexts that are closely related to the urban setting, like airports and banks (Simonite, 2021).

The risk of enhancing surveillance and undermining civil liberties

There are two principal types of surveillance-focused risks associated with adopting smart city technologies. First, when smart city technologies are deployed and adopted in the policing or law enforcement context they are, almost by definition, tools for enhanced surveillance which are deployed with the objective of enhancing public security or public health. The risk of enhancing surveillance through digital technologies is that doing so comes at too high a price for civil liberties. What that balance is, is a policy and political question that is context-based. The second form is the risk that almost all forms of smart city technologies—so long as they collect information that can be traced back to a particular individual or community—are, potentially, tools of surveillance without that being part of their initial purpose (see the discussion on mission creep and function creep in the discussion on surveillance).

The issue with enhancing surveillance in the name of public security

Surveillance is, most often, an affectation of privacy rights, which is commonly accepted as an important right to sustain other rights and liberties, like freedom of speech and association or freedom of movement (Solove, 2007). In all jurisdictions, however, certain forms of surveillance are accepted and justified as important to protect public values, like public

security and health. The general standard is that the erosion of civil liberties that results from surveillance practices should be proportional to the goal pursued (European Court of Justice, 2014; Solove, 2007). What is considered proportionality, however, may vary in different jurisdictions as it depends on cultural and institutional preferences.

The risk of enhancing surveillance disproportionately to the goals pursued comes from the vast scale of the information that can be amassed and the ability of integrating different datasets and the capabilities these technologies can create to enhance authoritarian and discriminatory practices. Sara Brayne describes how these tools are adopted and used in "Predict and Surveil: Data Discretion, and the Future of Policing," when discussing the tools and services designed and sold by Palantir, a CIA-financed surveillance software company. Before, she explains, officers and analysts conducted mostly one-off searches in siloed systems regarding license plate data, interview (FI) cards, traffic citations, gang system, etc. The Palantir platform helps overcome this fragmentation by integrating previously disparate data sources into a single search, which allows officers to "drill down" much deeper on any one individual, address, car, or entity by accessing more data points collected from more disparate sources, all searchable in relation to one another. "Seeing the data all together is its own kind of data" (Brayne, 2020).

The counterargument is typified by what Daniel Solove has described, and criticized, as the "nothing to hide" argument. As the argument goes, only those who are engaged in illegal activities have a reason to hide this information. There may be some cases in which the information might be sensitive or embarrassing to law-abiding citizens, the limited disclosure lessens the threat to privacy. Moreover, the argument goes, the security interest in detecting, investigating, and preventing crime is very high and outweighs whatever minimal or moderate privacy interests law-abiding citizens may have in information gathered by these kinds of surveillance tools and programs (Solove, 2007).

As Solove has contended, however, the "nothing to hide" argument stems from the faulty premise that privacy is about hiding a wrong. There are harms that can occur with vast surveillance programs that can be best described as chilling social beneficial behavior, like free speech and association, or leading to important power imbalances that adversely affect social structure, like excessive police or executive power (Solove, 2007).

Additional kinds of potential harms created by enhanced surveillance can be described as ones related to the information processing practices that occur when these tools are deployed at vast scale. One is, for example, increasing an individual's vulnerability to potential abuse of their information, or, simply, leading and strengthening systems that make decisions about individuals without their knowledge and participation (Solove, 2007; Véliz 2021).

Another type of potential harm is related to the structural biases and inaccuracies that can be embedded in these vast data sets. Along the lines of the risks of enhanced inequality and discrimination, critics have pointed out that metaphors about the fluidity or that describe data as a resource obscure the social, contextual, and political nature of data and the processes by which data comes into existence in the first place. In fact, data use and integration are nowhere near seamless and data sources are most often a patchwork of legacy systems, brought about at different times, used by different people, and often unable to communicate with each other (Fischer & Streinz, 2021; Boyd & Crawford, 2011). Thus, for example, the input and outputs of searches never arrive in the world formed. They are the result of different decisions and contexts by individuals. The inputs and outputs of the big data queries that feed these systems are often based on the imagined perceptions and

judgments of the operator about where a crime is more likely to happen or the kind of features a suspect may have (Boyd & Crawford, 2011; Brayne, 2020).

Lastly there is the risk of unchecked power. Discussing the United States Surveillance program, Daniel Solove argues that "[t]he issue . . . is not whether the NSA or other government agencies should be allowed to engage in particular forms of information gathering; rather, it is what kinds of oversight and accountability we want in place when the government engages in searches and seizures" (Solove, 2007). More than arguing that all forms of surveillance, or even enhanced surveillance, is harmful, per se, the goals of security are indeed important and their potential to address them real. The key question is about the mechanisms of oversight in place to mitigate some of the above-mentioned risks (see the chapter on accountability in this volume).

The European Court of Justice has reached a similar conclusion. The Court has examined a few times the legality of mass surveillance programs concerning the retention of data relating to electronic communications "to prevent, detect and investigate and prosecute crime and safeguard the security of the State" (European Court of Justice, 2014, 2020). It has found that this kind capability can be justified as it facilitates the investigation, detection, and prosecution of serious crime (European Court of Justice, 2020). Yet, it has also stated that it infringes upon the privacy rights of citizens and represents a risk mainly for individuals whose conduct in no way justifies the retention and analysis of data relating to them. It exposes individuals to a greater risk that authorities conduct investigations about them (European Court of Justice, 2014; European Court of Justice, 2020). To balance these competing interests, the Court has stated several times that the surveillance programs must be those only necessary and appropriate to achieve the public interest goals, which requires laying down clear and precise rules governing the applications of these technologies to personal data and imposing safeguards so that persons whose data have been retained have sufficient guarantees to effectively protect their personal data against the risk of abuse and against any unlawful access and use of that data. (European Court of Justice, 2020).

The risk of enhancing surveillance through tools that are not meant for surveillance in the first place

The risk of enhancing surveillance through tools that are not meant for surveillance in the first place refers to the fact that even when certain digital technologies are not meant to surveil citizens, they may hold the capabilities to do so. An example of the risk of smart city technologies not originally intended to provide security services is a 2020 case in San Diego, California, in the United States. In 2016, San Diego launched a smart streetlight pilot with General Electric that would save the city money and energy and would collect air quality and mobility data. Smart streetlights are, also, a rather normal fixture of the international "smart city" landscape. The program was extended full-fledged in 2018 and 14,000 of the city's 60,000 streetlights were replaced with the smart streetlights.

As it turns out, however, the city's police department had been regularly using the cameras in the streetlights since early 2019. Although the police did not have full access to the footage, and even wrote its own policy governing the use of street light data, at some point in 2019 it changed its policy from having to file a written request to access a portion of video to engaging its own video management service that gave them access within hours. In 2020, a local public scandal erupted when it came to public light that the local police

department had turned to the streetlights to surveil Black Lives Matter protesters in 2020. In its defense, the local police argued that footage from the streetlights has been used both to convict and exonerate suspects and that, in the context of the Black Lives Matter protests, the footage requested was only to follow up on looters and "civil unrest." At the same time, local advocacy groups and some members of the city council have said that the concern remains that the data would be misused (Holder, 2020). This example raises the question of whether all, or many, smart city technologies can potentially become tools for enhanced surveillance (Williams, 2021).

As a response to this event, San Diego's City Council passed an ordinance that now requires city staff to submit written reports on the proposed uses of potential surveillance technologies and analyze their impacts. In doing so, San Diego joined a small group of about 15 cities in the US that, as of mid-2020, had passed local surveillance laws, and created some opportunities for oversight before surveillance occurs and accountability mechanisms for law enforcement to report how it is using surveillance technologies (Fidler, 2020).

The risk of overrelying on smart city technologies to solve complex issues

A last, important critique to the deployment of smart city technologies is that they are often adopted in replacement of other, more integral policies or solutions to the issue the technology intends to solve. This is what information scholars call "technological solutionism," the indiscriminate use of technologies to solve complex problems, overestimating the potential of technology-driven solutions to do so (Green, 2019; Morozov, 2013). As this critique goes, conceptualizing city issues as complex as security and enhancing quality of life as problems that can be optimized by collecting and analyzing vast troves of data, risks simplifying complex political questions and undermining local democracy. The critique points out that making cities better in the name of economic growth, welfare, or safety through technology is a form of techno-utopianism; a way of reaching out to technology as if technology was a panacea able to solve complex problems alone (Baykurt, 2020). This creates incentives for policy makers to "attack the symptoms" but not the more structural causes of the problem (Moore Gerety, 2021).

One of the best examples of this line of critique is Ben Green's book *The Smart Enough City*. In it, Green shows that many smart-city solutions don't draw from a clear policy plan or a research process that is focused on the problem at hand but, rather, on how to implement a particular kind of technology. He thus describes many of these solutions as a "myopic reconceptualization of cities into technology problems" (Green, 2019). Smart city technologists assume that it is clear what services need to be delivered and how, ultimately leading to narrow conceptions of the problem at issue, and also to narrow, often ineffective, solutions. At the same time, Green shows that technological solutions embed certain values, most notably, the values of designers and the corporate actors behind them. Thus, for example, when cars destabilized cities' streets and cities stakeholders jockeyed to define how cars should be used, cities turned to engineers for a solution—engineers had displayed technical expertise in helping cities efficiently manage overburdened public utilities, such as water and electricity. The problem was that engineers understood the problem as one of managing traffic efficiently and, consequently did so at the cost of the needs of pedestrians and they helped "redefine streets as motor thoroughfares where pedestrians did not belong" (Green, 2019).

A similar example of this problem can be drawn from Chicago's heat list. The list allows officials to map social networks and identify youth at risks that could, potentially, lead to interesting approaches that treat violence like a public health risk and better understand systemic issues. However, as other scholars studying the list have pointed out, studying the data to forecast who might be engaged in violence does not automatically end the violence (Fergusson, 2017). Despite the heat-lists accuracy, questions remain about the program effectiveness and whether enough has been done to remedy the social and economic risks identified and, thus, its overall effectiveness (Fergusson, 2017).

A Political Economy Approach of Smart Cities and Smart Policing Tools

Political economy is a method of analysis that studies the social relations of (often loose) groups with power (Kennedy, 2019). The elements of a political economy of technology are the institutions, the market, and the ideology. Institutions are the explicit or implicit rules in a relationship between actors. Social norms are systems of institutions. Law, for example, is a particular institution that constrains and gives affordances to different actors in an explicit way. The market is another kind of institution. It is largely made of the constraints afforded by law (especially property and contract law) and social norms, and the complex collective interactions of actors within those structures (Kennedy, 2018) The market thus creates constraints, like price and competition, which make it easier for some actors, like corporations, to prevail over others (Lessig, 1999). Ideology is the prevailing set of ideas about how the world works; that is, how we understand somethings to cause others. Ideology is a site of struggle to shape these understandings and what is considered possible (Benkler, 2019). This political economy of technology framework draws mostly from the work of Yochai Benkler (Benkler, 2019).

This section expands on how the institutional and ideological context in which many of these technologies are designed and adopted has an important role to play in shaping the distributive effects of smart city technologies and, potentially the occurrence of their harmful. It focuses mostly on the role of formal constrains, like law. To do so, it starts with a general example on the role of law and fiscal incentives in the transformation of cities to accommodate cars. Then, it makes a deeper dive into the institutional and ideological factors that have shaped the adoption of smart-policing tools in the US and their deployment.

A first, general example: How law contributed to an old tech-driven transformation of cities

Take again one of the largest technologically driven urban transformations in history: the advent of cars. In Ben Green's technological solutionism critique, a crucial driver of this transformation in the United States was the role of engineers and designers who understood the problem as one of managing traffic efficiently. They did so by prioritizing cars over pedestrians (Green, 2019). There was an important role, however, for other institutional

arrangements like law and other incentives that placed engineers in that position to the first place.

When cars first arrived in cities in the United States, in the 1910s and 1920s, the reaction in cities was one of fear and resistance and some local laws made it hard to use cars by, for example, requiring motorists to advertise their intention of going upon the road one week in advance or to hire a person to walk ahead of the car, bearing a red flag (Jeansone, 1974). Similarly, until the 1900s, street parking was broadly outlawed. In the coming years, however, the legal system and vast funds were mobilized at large to finance and accommodate streets, roads, and highways, and to facilitate changes in the spatial arrangement of residences, commerce, and industries toward sprawl (Shill, 2020).

The New Deal, for example, put huge sums of money to work on building new roads and highways (Gutfreund, 2005). In cities, land use laws that limited denser development were crucial to pave the road for car-centric sprawled metropolitan areas (Shill, 2020). Real estate development and regulations also contributed to incentivize the building of office buildings, shopping centers, and more residential areas to provide amenities for the suburbs (Gutfreund, 2005). Thus, at the time, automobiles allowed people and firms to cover further distances, and policies oriented towards favoring urban sprawl and invigorating the economy by building a particular type of infrastructure contributed to a particular transformation of the American urban landscape (Shill, 2020). Considering these institutional factors shows how policies and laws orient city officials and planners to adopt these technologies in certain ways and not in others. It turns out that federal government and private grants that fund smart-city technologies and local government structures may be pushing cities towards partnering with transnational companies to upgrade their infrastructures, attract different forms of investment, and create employment and business opportunities from which they can then collect revenue.

A second, contemporary example: The adoption of big data policing tools

How policies and laws orient city officials and planners to adopt these technologies in certain ways is also observable upon a closer look at the adoption of data-driven technologies. Take as an example the adoption of big data policing tools in the United States.

In this context, three main factors seem to have come together to shape a particular form of data-driven policing. First, was the widespread use of the internet, and the fall in costs of data storage and processing. The second element is ideology. Many studies have shown that public security related policies—from policing to the adoption of surveillance technologies—are often influenced by racism and fear of minorities and have driven police and criminal law enforcement policy (Benjamin, 2019). Lastly, law and the grants and fiscal incentives local governments largely adopted after 9/11 have been very influential towards the adoption of these tools, too. Indeed, after 9/11 there was a move towards anti-terrorism practices in local police departments that seems to have institutionalized predictive policing and that gave local police departments access to grants to buy big data technologies. Local budget cuts that after the financial crisis of 2008 led local police agencies to dispense of an important force of their workforce and, importantly, some of their accompanying

services, like social services. Thus, although police budgets seem to have grown in net, the federal grants that were part of the post-9/11 package became their main form of funding and, the resulting technologies became their main form of working. If so, and as the old idiom goes, your only tool is a hammer, then every problem looks like a nail. To show how this might have happened, what follows briefly explains how policing services are typically funded and how this changed after 9/11.

City governments are, in general, often asked to do a lot with a little. Municipalities, to begin with, have limited powers. Local government law is the body of law that grants cities' formal authority to engage in certain activities—to decide to enforce a minimum wage or to provide certain social services over others. Cities are not only bound by law but, of course, by their financial constraints. This discussion focuses on how financial constraints and incentives have incentivized city police departments to adopt predictive policing tools.

At least since the 1960s, local government law in the United States has been largely shaped to make municipalities principally dependent financially on the taxes they can levy within their territories—this power, too, shaped by law. So-called market-based local government law intends to enforce some form of financial soundness in public spending and is animated by the economic truth that capital and labor are more mobile than local governments and cities, which are fixed by nature (Schragger, 2016). Consequently, localities are advised to attract individuals and firms through the different "packages" of services and taxes they offer—perhaps higher property taxes and good schools in wealthy suburbs, or low property taxes, but higher income taxes and better transportation systems in service-industry oriented cities. Market-based local government law also typically advises cities not to play an active role in fulfilling the redistributive functions of government: not because they aren't important, but because residents and firms that bear the burden of local redistribution can easily exit to neighboring jurisdictions that don't impose those burdens on them. Thus, redistributive policies should be left to higher spheres of government (Gillette, 2007).

Another important source of income for cities is federal and state aid. Starting with the Reagan administration, however, overall federal aid to cities has decreased in response to persistent budget deficits and a resurgent believe in devolution. Although state aid has increased since again, this resulted in cities having to downsize many of their social services like housing, community development, and urban economic development. As of the early 2000s, K–12 education had become the leading beneficiary of state aid (Wallin, 2005).

Despite this general decline, federal grants from the 1994 Violent Crime Control and Law Enforcement Act (also known as the Clinton crime bill) and emergency preparedness funding from the U.S. Department for Homeland Security (DHS) seem to have kept city police departments afloat. An important element of police spending in the 1990s was the Community Oriented Policing Services program (COPS), a program of federal aid that helps fund additional police officers and gave increased attention to anti-crime efforts in cities (U.S. Department of Justice, 2014. After the 9/11 attacks, the DHS increased federal aid to local police departments for "emergency preparedness" ten-fold (Homeland Security, 2008).

Indeed, the 9/11 terrorist attacks on the United State unleashed an important transformation, or evolution, in policing techniques. Already since the 1960s and 1970s city planners, and city planning research institutions funded by the federal government, had started to use management techniques developed in World War II that heavily integrated data analytics (Williams, 2020. In the 1990s, a community policing model had emerged, which focused

on preventing crime by addressing its underlying conditions and delegating some decision making to officers through a problem-solving approach. After 9/11, intelligence-led policing emerged as an important compliment to community-based policing but extended it to research-based approaches, powered by information and communications technologies which emphasized the role of data analytics to "quickly identify and focus on crime hot spots, enabling police practitioners to respond rapidly and to counter crime problems successfully" (Beck & McCue, 2010).

In 2009, Congress enacted $3 billion to support state and local assistance programs to prepare them to "respond to threats or incidents of terrorism and other catastrophic events" (U.S. Government Printing Office, 2009). The grants were part of the "Implementing Recommendations of the 9/11 Commission Act of 2007" which implemented some of the recommendations of the 9/11 Commission, importantly a new method of redistributing anti-terrorism funding, to assist "[s]tate, local, and tribal governments in "preventing, preparing for, protecting against, and responding to acts of terrorism" (U.S. Congress, 2007. The Act also established an Urban Area Security Initiative to provide grants to assist high-risk urban areas. High-risk urban areas were defined on grounds such as its population, including appropriate consideration of military, tourist, and commuter populations, its population density, history of threats, etc. (U.S. Congress, 2007). These grants were also intended to assist these areas in achieving "target capabilities related to preventing, preparing for, protecting against, and responding to acts of terrorism," which included "protecting a system or asset included on the prioritized critical infrastructure list; ""purchasing, upgrading, storing, or maintaining equipment, including computer hardware and software;" "ensuring operability and achieving interoperability of emergency communications" "establishing, enhancing, and staffing with appropriately qualified personnel state, local, and regional fusion centers; and purchasing various other equipment and devices to enhance their IT public safety networks". (U.S. Congress, 2007. No more than 50 percent of the amount awarded to a city could be used to pay for personnel or to construct buildings or other physical facilities (U.S. Congress, 2007).

After the 2008 financial crisis, city departments experienced important financial cuts from their local sources, policing being no exception. According to a 2013 report of the Police Executive Research Forum (PERF), most police departments and agencies experienced important budget and cutbacks during this time. Recruiting, training, officer pay raises, and reductions in police services were amongst the main areas that experienced cuts (Police Executive Research Forum, 2013). As a result, PERF documented how "smart policing" strategies were adopted by different departments. These strategies often involved eliminating or reducing specialized units, instead of reassigning all available and sworn officers to patrolling the streets (in line with the community policing ideas adopted around the time of the COPS program). At the same time Real Time Operation Centers (RTOC)—and personnel assigned to the RTOC—monitor the incoming data and quickly disseminate information to the officers on the street who are patrolling. It was, however, while city governments were cutting the workforce in their police department that the DHS grant program was going full-fledged. In fact, the DHS's Urban Areas Security Initiative slightly grew from 2009 to 2010 (U.S. Department of Homeland Security, 2009).

In in this context, Charlie Beck, the former chief of the Los Angeles Police Department wrote a seminal article on fighting crime in a recession in which he highlighted the importance of anticipating or predicting future demand for policing services. Using an already

famous example, Beck pointed out that just as Walmart was able to shift its supply chain to anticipate consumption of Pop-Tarts in association with weather events, "the discovery part of predictive analysis . . . can be tremendously powerful in policing" (Beck & McCue, 2010). The main advantage from these tools was coming also from the institutional context: police departments were increasingly experiencing severe budget cuts and predictive policing—and its emphasis on prevention, targeted response, and resource allocation—supported the idea of doing more with less (Beck & McCue, 2010). At the same time, however, the funding structure for local police departments was such that data-driven and preemptive strategies were favored over recruiting, training, and reducing police services and "non-sworn" personnel were importantly cut from police departments.

Conclusion

This chapter was intended to familiarize readers with the main narratives and risks of smart city technologies and the political economy elements that shape their adoption. As a compliment to the main narratives about smart city technologies, the chapter showed how political economy elements shape the adoption and diffusion patterns of these tools. The chapter showed that even though digital technologies deployed in the city context hold the promise of making all sorts of processes, planning, and services more efficient, seamless, and quick, they can also entrench discrimination, inequality, and surveillance in a city's fabric, and these effects cannot be easily displaced from the social and institutional contexts in which they operate. Many tools—like energy consumption reduction AI systems for buildings, for example—operate in institutional contexts that are maybe not particularly problematic, from an equality and fairness perspective. Others, like the surveillance tools that we mostly discuss in this chapter, do raise some questions and present important tradeoffs to policymakers and societies at large. The chapter focused on presenting the risks and their relation to political economy factors, with the hope of informing the decisions and judgment of actors interested in the governance of these tools.

I have suggested that the role of technologies and how they are designed, diffused, and deployed depend importantly on political economy elements, like preexisting institutional arrangements, the market, ideology, and social shocks. This chapter did this by diving deeper on the factors that have shaped the adoption and regulation of predictive policing tools in the US in the past 20 years—like the 9/11 terrorists attacks, the 2008 financial crisis, and the Black Lives Matter movement. Their distributive effects and how they impact people's lives can be better understood and tackled by taking those elements into account. Tentatively I also suggested that the rise of big data policing in the US, with a focus on preventing criminal acts but with a lack of focus on tackling the complex underlying social and economic causes that lead to crime, may have been influenced by DHS grants that, post 9/11, facilitated local access to these technologies. At the same time, following the financial crisis, local police departments experienced a cut in personnel that made addressing and understanding underlying causes to crime and violence less of a priority. Recall, however, that after the BLM protests in the years leading to 2020, local social movements may be able to impose some accountability mechanisms on police departments adopting these technologies. This may lead, at least, to controlling their risks. This suggests there is room

for mass-digitization that is not mass-surveillance, and that there is potential to disrupt, review, and carefully consider how big data tools are used in these sensitive contexts.

This is a tentative analysis and hypothesis that most likely presents important variations in different cities, with different power and budgets, and where different cultural preferences and pressures are prevalent. Understanding those particularities may be ideal case study for future researchers.

References

Ash Center for Democratic Governance and Innovation—Harvard Kennedy School. (2017). Our mission. Data smart city solutions. https://datasmart.ash.harvard.edu/.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*, June 23. https://www.wired.com/2008/06/pb-theory/.

Atehortua, J. (2021). Presentation at Session 4: ¿Medellín nos cuida? Inteligencia artificial y big data para la seguridad. Edgelands Institute. https://www.youtube.com/watch?v=5Py4elVIUk&t=3s.

Baykurt, B. (2020). Can smart cities be equitable cities. *Public Books*. https://www.publicbooks.org/can-smart-cities-be-equitable-cities/ [https://perma.cc/5CWL-93X8].

Baykurt, B., & Raetzsch, C. (2020). What smartness does in the smart city: From visions to policy (Vol. Series 10). University of Massachusetts Amherst, Communication Department Faculty Publication. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1113&context=communication_faculty_pubs [https://perma.cc/T9LV-Z32K].

Beck, C., & McCue, C. (2010). Predictive Policing: What can we lLearn from Walmart and Amazon about fighting crime in a recession. *Police Chief Magazine*. https://www.policechiefmagazine.org/predictive-policing-what-can-we-learn-from-wal-mart-and-amazon-about-fighting-crime-in-a-recession/.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Polity.

Benkler, Y. (2019). Power and productivity: Institutions, ideology, and technology in political economy. *SSRN3*. https://ssrn.com/abstract=3503962.

Botero Arcila, B. (2021). The place of local government law in the urban digital age. *SSRN3*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3848202.

Boyd, Danah and Crawford, Kate, Six Provocations for Big Data (September 21, 2011). A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011, Available at SSRN: https://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431

Brayne, S. (2020). *Predict and surveil:Data, discretion, and the future of policing*. Oxford University Press.

Centre for Internet & Human Rights at European University Vladrina. (2015). The ethics of algorithms: From radical content to self-driving cars. Centre for Internet and Human Rights. https://cihr.eu/wp-content/uploads/2015/07/Ethics_of_Algorithms_Final_Paper.pdf.

Chicago Police Department. (2015). Custom Notifications in Chicago, Special Order S10–05 III.C. October 6. http://directives.chicagopolice.org.

Cisco. (n.a.). What is a smart city? *Cisco Cities and Communities*. https://www.cisco.com/c/en/us/solutions/industries/smart-connected-communities/what-is-a-smart-city.html.

Davey, M. (2016). Chicago police try to predict who may shoot or be shot. *New York Times*, May 24. https://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html.

European Commission. (2021). Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. E.C. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Court of Justice. (2020). Judgments in Case C-623/17, Privacy International, and in Joined Cases C-511/18, La Quadrature du Net and Others, C-512/18, French Data Network and Others, and C-520/18, Ordre des barreaux francophones et germanophone and Others. ECJ.

European Court of Justice. (2014). Judgement in Case C 293/12, Digital Rights Ireland Ltd, and C 594/12, Kärtner Landesregierung and Mr Seitlinger, Mr. Tschohl and Others. ECJ.

Fergusson, A. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. NYU Press.

Fidler, M. (2020). Fourteen places have passed local surveillance laws. here's how they're doing. *Lawfare*. https://www.lawfareblog.com/fourteen-places-have-passed-local-surveillance-laws-heres-how-theyre-doing.

Fischer, A., & Streinz, T. (2021). Confronting data inequality. World Development Report 2021 background paper. IILJ Working Paper 2021/1.

Frug, G. (1980). The city as a legal concept. *Harvard Law Review 93*, 1057.

Gerety, R. (2021). Inside the rise of police department real-time crime centers. *MIT Technology Review*. https://www.technologyreview.com/2021/04/19/1022893/police-surveillance-tactics-cameras-rtcc/.

Gillette, C. P. (2007). Local redistribution, living wage ordinances, and judicial intervention. *Northwestern University Law Review 101*(3), 1057.

Goldsmith S., & Crawford, S. (2014). *The responsive city: Engaging communities through data-smart governance*. John Wiley & Sons.

Green. B. (2019). *The smart enough city: Putting technology in its place to reclaim our urban future*. MIT Press.

Grother P., Hom, A., Ngan, M., & Hanaoka, K. (2021). Face recognition vendor test: Identification for paperless travel and immigration. National Institute of Standards and Technology. U.S. Department of Commerce. NIST. Report Number 8381. https://www.nist.gov/publications/face-recognition-vendor-test-frvt-part-7-identification-paperless-travel-and.

Gufflet, B., & Kremp, D. (2021). Across the Blocks: A Journey Into Smart Cities. https://www.acrosstheblocks.com/the-study.

Gutfreund O. (2005). *20th-century sprawl: Highways and the reshaping of the American landscape*. Oxford University Press.

Hill, K. (2021). Your face is not your own. *New York Times*, March 18. https://www.nytimes.com/interactive/2021/03/18/magazine/facial-recognition-clearview-ai.html.

Holder, S. (2020). In San Diego, "smart" streetlights spark surveillance reform. *Bloomberg.com*, August 6.https://www.bloomberg.com/news/articles/2020-08-06/a-surveillance-standoff-over-smart-streetlights.

IBM. (n.a.). IBM smarter cities—future cities—United States. IBM. http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/.

Jeansone, Glen (1974). The Automobile and American Morality. *Journal of Popular Culture Vol. 8, Issue 125*.

Kennedy, D. (2018). *The stakes of law, or Hale and Foucault!* Harvard Law School, Legal Studies Forum.

Kennedy, D. (2019). Transformations of the law of the globalizing economy. In P. Kjaer (Ed.), *The law of political economy: Transformation in the function of law* (pp. 125–202). Cambridge University Press.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society* April–June, 1–12.

Lessig, L. (1999). The law of the horse: What cyberlaw might teach. *Harvard Law Review 113*(2), 501–549.

Mattern, S. (2021). *A city is not a computer: Other urban intelligences*. Princeton University Press

McKinsey Global Institute, Woetzel, J., Remes, J., Boland, B., Lv, K., Sinha, S., Strube, G., Means, J., Law, J., Cadena, A., & Tann, V. V. D. (2018). Smart cities: Digital solutions for a more livable future. McKinsey & Company. https://www.McKinsey.com/~/media/McKinsey/industries/capital%20projects%20and%20infrastructure/our%20insights/smart%20cities%20digital%20solutions%20for%20a%20more%20livable%20future/mgi-smart-cities-full-report.pdf.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.

O'Neill C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishers.

Police Executive Research Forum. (2013). *Policing and the economic downturn: Striving for efficiency is the new normal*. Critical Issues in Policing Series.

Porter, J. (2019). Federal study of top facial recognition algorithms finds "empirical evidence" of bias. *The Verge*, December 20. https://www.theverge.com/2019/12/20/21031255/facial-recognition-algorithm-bias-gender-race-age-federal-nest-investigation-analysis-amazon.

Rao, L. (2010). The Final Tally: More Than 1100 Cities Apply for Google's Fiber Network, *Tech Crunch*. https://techcrunch.com/2010/03/27/the-final-tally-more-than-1100-cities-apply-for-googles-fiber-network/?guccounter=1.

Rhee, N. (2016). Can police big data stop Chicago's spike in crime? *The Christian Science Monitor* June 2. https://www.csmonitor.com/USA/Justice/2016/0602/Can-police-big-data-stop-Chicago-s-spike-in-crime.

Shill, G. (2020). Should law subsidize driving? *NYU Law Review 95*, 499.

Schragger, R. C. (2016). *City power: Urban governance in a global age*. Oxford University Press.

Simonite, T. (2021). Face recognition is being banned—but it's still everywhere. *Wired*, December 21. https://www.wired.com/story/face-recognition-banned-but-everywhere/.

Solove, D. (2007). "I've Got Nothing to Hide" and Other Misunderstandings of Privacy. *San Diego Law Review 44*, 745.

U.S. Congress, Implementing Recommendations of the 9/11 Commission Act of 2007. Public Law 110-53. https://www.congress.gov/110/plaws/publ53/PLAW-110publ53.htm.

U.S. Department of Justice. (2014). The Cops Office: 20 Years of Community Oriented Policing. https://cops.usdoj.gov/RIC/Publications/cops-p301-pub.pdf

U.S. Department of Transportation. (2017). Smart city challenge. https://www.transportation.gov/smartcity.

U.S, Executive Office of the President. (2016). Artificial intelligence, automation, and the economy

U.S. Government Printing Office, 3. Homeland Security Funding Analysis (2009). https://www.govinfo.gov/content/pkg/BUDGET-2009-PER/html/BUDGET-2009-PER-4-1.htm.

U.S. Homeland Security. (2008). Department of Homeland Security annual financial report. DHS. https://www.dhs.gov/sites/default/files/publications/cfo_afrfy2008_0.pdf

U.S. Homeland Security. (2009). FY 2010 Preparedness Grant Programs Overview. https://www.dhs.gov/xlibrary/assets/grant-program-overview-fy2010.pdf

Véliz, C. (2021). If AI is predicting your future, are you still free? *Wired*, December 27. https://www.wired.com/story/algorithmic-prophecies-undermine-free-will/.

Wallin, B. (2005). Budgeting for basics: The changing landscape of city finances. *Brookings*. https://www.brookings.edu/wp-content/uploads/2016/06/20050823_BudgetingBasics.pdf.

Williams, R. (2021). Whose streets? Our streets! (Tech Edition). *2020–21 "Smart City" Cautionary Trends & 10 Calls to Action to Protect and Promote Democracy*. https://whosestreets.substack.com/p/finalreport.

Williams, S. (2020). *Data action: Using data for public good*. MIT Press.

9/11 Commission Act of 2007. Retrieved from https://www.congress.gov/110/plaws/publ53/PLAW-110publ53.pdf Sec 2008 (b)(2)(B)

CHAPTER 41

ARTIFICIAL INTELLIGENCE IN HEALTHCARE

NAKUL AGGARWAL, MICHAEL E. MATHENY, CARMEL SHACHAR, SAMANTHA WANG, AND SONOO THADANEY-ISRANI

Introduction

Artificial intelligence (AI) in healthcare is witnessing a surge of AI-driven innovations that could potentially transform the current landscape, from administrative optimization to individual diagnosis and treatment to public health monitoring. While AI holds great promise for improving patient outcomes and health system efficiency, it is crucial that governmental agencies, industry partners, academia, and the healthcare ecosystem cooperatively and deliberately ensure AI is implemented equitably. The prioritization of regulatory guidelines supplemented by best practice certification or accreditation is critical to promote fair and effective use and foster trust and wider adoption (Roski et al., 2021).

The future of AI in healthcare rests on the continual evolution of AI and its integration into health systems, while designing and evaluating interventions based on the Quintuple Aim, a framework geared towards optimizing health care delivery (National Academy of Medicine, 2019). It centers on patient outcomes, cost reduction, population impact, clinician wellness, and equity and inclusion. Techno-chauvinism must be avoided. Often, technologists build novel tools and seek problems to solve with them. However, implementation science guides us to a user and use-case centered process solving real-world problems, for real-world people.

Finally, governance and decision-making should be guided by a range of relevant frontline stakeholders—patients, caregivers, clinical teams, community-based health care organizations, and diversity of thought, gender, race, education, culture, and more should be prioritized. Else, we risk automating poverty and adding to the existing structural determinants of health anchored in the inequities of colonialism, capitalism, sexism, homophobia, and various forms of supremacy, including but not limited to White and Male. Here, in the context of the goals of the Quintuple Aim, we provide (1) a critical survey of state-of-the-art clinical AI tools and promising future directions; (2) an overview of the processes underlying clinical AI algorithm development, organizational implementation,

and provider adoption; (3) an analysis of how bias can emerge in and perpetuate via healthcare AI tools and the necessary steps to avoid it; and (4) an overview of national and international frameworks for the regulation of AI-driven tools in healthcare.

A Brief Survey of Clinical AI Applications

AI can be a powerful tool in predicting and prognosticating medical conditions and in aiding medical decision-making, and it has made iterative gains in accuracy and capacity in recent years. Deep learning techniques have been shown to outperform human subspecialists or conventional methods in the diagnosis of melanomas (Maron et al., 2019), bladder cancers (Vriesema et al., 2000), structural heart disease, and excessive fetal weight gain.

Historically, the wide use of algorithms in healthcare can be traced to the field of cardiology, which relied on risk assessment calculators to diagnosis and treat patients at highest risk for cardiovascular disease. One of the first risk calculators developed was the Framingham 10-year Risk Score in the late 1990s, which estimated one's 10-year risk of developing coronary heart disease (CHD). The data was largely derived from the Framingham Heart Study, a seminal research cohort of its time, but made up of notably predominantly White individuals of Western European descent (Tsao & Vasan, 2015). By 2013, an alternate 10-year risk equation was developed and recommended by the American College of Cardiology and the American Heart Association: Atherosclerotic Cardiovascular Disease (ASCVD Risk Estimator), with data derived from a cohort of predominantly non-Hispanic White or Black populations. This equation is now ubiquitously used. However, the general population of the United States is heterogenous, leading to miscalculation of risk in under-represented racial and ethnic minorities. Traditional risk calculators have been criticized for being inaccurate in these populations (D'Agostino et al., 2001) and too reliant on assumptions of statistical linearity (Gijsberts et al., 2015).

This is an area of opportunity for machine learning (ML) models, which may use identical input variables but because of greater statistical versatility, flexibility, and the ability to learn intrinsic properties and patterns from a data set are able to arrive at more accurate predictions. In one example of a supervised ML model that used the same risk factors as that of the ASCVD Risk Estimator, the ML algorithm was able to detect 13 percent more high-risk individuals and reduce unnecessary drug therapy (i.e., over-prescription) by 25 percent in low-risk individuals compared to the ASCVD Risk Estimator in a multi-ethnic American population of White, Black, Hispanic, and Chinese patients (Kakadiaris et al., 2018). The algorithm was then externally validated in an external cohort of a non-multi-ethnic Belgium population and proved to be more accurate than the ASCVD Risk Calculator. As more data is introduced into risk calculators, the prediction of short-term cardiovascular events may become a reality, allowing clinicians to counsel patients and deter more imminent health risks.

The above narrative generalizes and reflects the larger trend of healthcare AI's gains in performance through deep learning technologies. Perhaps the most known among these are the image-based diagnostic tools, which have become the "poster child" for AI success in healthcare. One popular example has been early cancer detection using image interpretation support (Pesapane et al., 2018). For stand-alone computer interpretation of medical images to be acceptable, an algorithm's performance must exceed that of human clinicians. The accuracy of

AI systems in breast cancer detection has been shown to be superior to most radiologists but may under-perform when compared to the *best* radiologists (McKinney et al., 2020).

The recent explosion of mobile health and wearable devices marketed directly to consumers continuously collect health-related behavioral information and physiological measures, contributing to an improved patient experience and opportunities to advance population health initiatives. These are in sharp contrast to the current healthcare paradigm, where data is only captured at discrete provider visits. Such tools can be used for disease screening (Seneviratne et al., 2018), diagnosis (Yin et al., 2020), monitoring and management (Quinn et al., 2008), and for improving access to care (Correia et al., 2008), while empowering patients to access and monitor their own health data. A prominent example includes wearable arrhythmia monitoring devices, such as the Apple Watch (Koshy et al., 2018; Turakhia et al., 2019). These technologies have also revolutionized clinical trial designs, leveraging these omnipresent devices for study enrollment targets in a matter of months—a feat that would ordinarily take years (Turakhia et al., 2019).

One growing area of concern with wearable devices is the need for regulatory oversight. Some of these devices seek Food and Drug Administration (FDA) clearance for some functions, while other functions on the same device may not be FDA cleared. As manufacturers are allowed to classify device functions as non-medical, they can enter the consumer retail space with little regulatory oversight but still relay health information that patients may then bring to the attention of their healthcare providers. There are numerous potential risks to this practice, including over-diagnosis of conditions and associated harms of over-treatment.

As technology continues to evolve, remote audio and video technologies may allow for contactless detection of medical conditions in public spaces, raising further considerations and questions around privacy and data security. However, keeping in mind the goal of equity and inclusion in the Quintuple Aim, these devices are more accessible in wealthy countries and used among people that are tech-proficient and have more resources, which can further exacerbate health inequalities.

A question remains yet to be answered. Given its higher efficiency and purported potential for higher sensitivity and accuracy in diagnoses, will AI algorithms ever fully replace clinical staff? Given current discrepancies in algorithmic performance in real-practice settings, this is unlikely. Instead, they can be leveraged to optimize clinical workflow or automate routine, time-consuming clinical tasks, potentially alleviating burnout associated with excessive clerical duties and creating space and time for more empathic, humanistic clinical encounters (Topol, 2019). Exceptions to this include low-resource health settings in rural areas and some small and middle-income countries where physical and human resources may be limited, presenting an opportunity for AI platforms to play a unique role in improving access to and efficiency of care.

Overview of the Lifecycle of AI in Healthcare and Development of Clinical AI Algorithms

There are four key conceptual phases in the implementation lifecycle of healthcare AI: (1) assessing organizational needs and currents tools available, (2) AI development, (3)

implementation of AI tools, and (4) AI tool performance surveillance and maintenance. In this section, we highlight the key challenges and barriers for sustained, accurate AI solutions in healthcare.

Phase 1: Organizational need, readiness for change, and current state

Because context, workflow nuances, and pertinent use case scenarios should drive implementation practices, the lifecycle of AI begins with a clear understanding of needs, the problem to be solved, and decisions on how AI *could or could not* be relevant; these considerations will lead to enhanced patient and provider experience with new AI tools, as well as more prudent resource management in relation to AI tools. Until very recently, this has been an ignored and underappreciated phase of the overall modeling lifecycle. There are many instances of tools and models developed without a clearly defined problem that is appropriate for an AI solution, or cases of AI deployment into organizations that are unprepared to manage safety, accuracy, and related costs. A number of frameworks, summarized by Fan and colleagues, also note key challenges in behavior intention and sustained use (Fan et al., 2020).

Educational efforts to increase familiarity and understanding of the key concepts, strengths, and limitations of AI tools are critical to understand and effectively leverage AI (Abdullah & Fakieh, 2020; McGrail et al., 2017). While there is general agreement in the potential utility of AI applications, most respondents of surveys indicate a lack of understanding of how such algorithms function (McCradden et al., 2020). In turn, this negatively impacts an organization's readiness for change using these technologies and limits the capacity for understanding how these tools can be effectively leveraged for healthcare needs. Addressing this gap falls largely on national governments and large national/international stakeholder organizations (Baharuden et al., 2019). Exploration and adoption of AI technologies has varied across disciplines based on core data elements and perceived need. Increasingly, healthcare organizations and AI developers are calling for wider participation in the conceptualization and expansion of tools to ensure broader AI (Solomon & Rudin, 2020).

Managing perceptions and fears about humans being replaced by AI cannot be ignored because surveys of healthcare worker perceptions frequently report feeling they could potentially be replaced (Abdullah & Fakieh, 2020; Laï et al., 2020). The healthcare AI industry has repeatedly emphasized a need for adapting, retooling, and adjusting the scope of human practice to manage clinical care in each sub-discipline. Ideally, this would include building AI to accomplish tasks for which it is well-suited and focusing more human time and effort to those tasks that humans are best suited to deliver (Johnson et al., 2021; Wall & Krummel, 2019).

The majority of the physician population have strong interest in the application of AI to clinical practice (Laï et al., 2020; Sarwar et al., 2019; Wadhwa et al., 2020). However, most of the general public remain skeptical of allowing AI to replace the physicians and nurses in the tasks they usually perform (Fenech & Buston, 2020). Artificial intelligence is unlikely to completely replace the clinician workforce, but engagement from frontline stakeholders is necessary to ensure that trust is built with patients. This begins with medical education.

Presently, digital skills and familiarity with AI is not routinely taught in medical school curriculum globally. Medical education reform that incorporates skills such as digital literacy, computational and statistical concepts, clinical informatics, and machine learning is needed to increase provider trust and buy-in with AI technologies (Paranjape et al., 2019).

Phase 2: AI development

The AI development phase begins with defining a target state, appropriate outcomes, and data sources; defining the accompanying changes in the workflows and relevant stakeholders; and then generating the initial AI model for the intended target.

Facilitating *trust in algorithms* is a key tenet of AI development and is distinct from trust and transparency in the process, use case, and outcome targets discussed above. This requires sufficient transparency in the AI development process at the design level, the data level, and the algorithmic level. It also translates into data transparency, which is the process of reporting key characteristics so that users understand the clinical, socioeconomic, racial, ethnic, and other characteristics of the data used to train the algorithms. For algorithmic transparency, this is the process of documentation of the design, key assumptions, meta-parameters used, and any parallel or accompanying algorithms or analyses that increase explainability. Equally important for trust in AI systems is the reporting on performance and robustness of the algorithms at the initial development stage and later during surveillance and updating. As always, ensuring fairness and representativeness in the context of use promotes trust (Asan et al., 2020).

There have been several best practices, checklists, and recommendations for developing AI models, such as MI-CLAIM (Norgeot et al., 2020), MINIMAR (Hernandez-Boussard et al., 2020), and SPIRIT-AI (Rivera et al., 2020). However, these practices are erratically implemented (Crowley et al., 2020). Thus, certification and accreditation have recently been proposed as complementary mechanisms to governmental and regulatory frameworks. This could allow flexibility and innovation in AI systems while promoting trust and safety (Babic et al., 2021).

Overall, here the challenge is to balance safety while navigating innovation and regulation, and prioritizing trust, transparency, and equity.

Phase 3: Organizational AI implementation

The challenges in organizational-AI-implementation in healthcare are contextual and use-case-dependent. Adopting and sustaining AI in practice requires clinician-users to embrace the usefulness of the AI, which while obvious, cannot be overstated (Petitgand et al., 2020). If the tool generates errors or creates additional cognitive load, even in a different area than the one it reduces, it may not be a net gain for clinician-users. Key foci are the usability of the AI, including data aggregation into clinically familiar frames, user-friendly data visualizations, and integration into the workflow at the right time, place, and context. A number of relevant frameworks are from Human-Computer Interaction and Implementation Science domains (Dopp et al., 2020). Importing these frameworks requires particular adaptations and extensions for AI and warrants additional study and characterization for successful

adoption. One Dutch study underscored some of these key features, citing expectations of added utility, hospital investment, and clinical business ownership as positive implementation factors. Undocumented processes, lack of stakeholder inclusion in the process, and lack of trust were cited as negative factors (Strohm et al., 2020).

Implementation requires data encryption and privacy practices to be maintained and vigilantly defended (Rastogi et al., 2015; Shi et al., 2020). However, there is a critical need for adequate organizational governance, from health IT to system leadership to clinical business owners (Sohn et al., 2017; Roski et al., 2018; Hunt & McKelvey, 2019; Reddy et al., 2019). Ensuring that human review and consideration is promoted and incorporated will help the sustained safe use of these technologies and ultimately bolster patient outcomes (Holzinger et al., 2019; Lee et al., 2019).

Finally, organizational policies and careful consideration for ethical technological and AI practices must always be front and center, especially during implementation (Mudgal & Das, 2020; Tzachor et al., 2020).

In summary, implementation of AI requires a focus on organizational culture, processes, implementation science, and workflow integration before technical features. Thus, much still needs to be learned about the best ways to successfully implement AI in the real-world.

Phase 4: AI performance surveillance, updating, and maintenance

Without pre-planned and sustained maintenance of implemented AI, model drift is a certainty. This occurs when the statistical association between features and target variables evolves, causing degradation of AI model performance over time. Methods and mechanisms for surveillance and maintenance of data and performance should be considered as part of the organizational lifecycle approach to managing AI tools and systems. AI solutions for healthcare should also promote clinically or operationally relevant cut-points or performance requirements that are incorporated into the algorithms, rather than only statistically defined performance metrics. Some methods incorporate this into the model development itself, such as online learning, but others provide mechanisms to do it separately.

In 2020, there were new tools focusing on the surveillance of performance, using metrics of discrimination and calibration over time, to detect when a method is failing and needs updating (Davis et al., 2019; Kelly et al., 2019; Subbaswamy & Saria, 2019). Once detected, there should be methods in place to help address this with automated frameworks to help scaling (Davis et al., 2019).

Finally, careful consideration must be given to model shift and dataset shifts over time. An example of model shift is the case of Google Flu Trends, which consistently underestimated annual influenza predictions due to shifts in search engine behaviors created by Google's own modifications to its search algorithms (Lazer et al., 2014). Similar use of open sources of information, such as Twitter, Facebook, and other internet-based services, must consider that user actions and behaviors change over time, and a previously fitted model may not be as accurate forever. Within healthcare, clinical guidelines and treatments are routinely changing and updates to the electronic health record (EHR) may change staff actions, such

that a previously fitted AI model may no longer be reliable. Mitigation efforts to identify drift are needed to ensure consistent performance over time.

Instead of a linear process, where implementation is viewed as the "final step" of an AI model, an iterative development process is necessary to ensure success of any clinical AI product. Successful adoption and integration of AI innovations require understanding and engaging stakeholders, along with their interdependence, workflows, and cultural dynamics.

Developing AI to be Fair, Equitable, and Inclusive

With the rapid innovation of AI healthcare tools, we must ensure that these systems have the intended outcomes for *all* target populations. AI solutions should not be viewed as neutral algorithms, but instead as products of human-design and judgment with downstream effects of informing, augmenting, or automating healthcare decisions. The Allegheny Family Screening Tool (AFST) (Chouldechova et al., 2018), a predictive risk model for child abuse, is a noteworthy example of the potentials of automating bias. First launched in 2016, the AFST tool provides child abuse hotline employees with a score between 1 to 20 to identify children at the highest risk of being removed from their homes within the next two years and could trigger a subsequent investigation. The algorithm is trained on data from public service records on welfare assistance, welfare assistance, drug and alcohol services, housing, and previous phone calls to the hotline. However, calls to the hotline from healthcare providers or community workers are three times more likely to report Black families than non-Black families and reflect implicit racial biases, which ultimately may reinforce misguided predictions and institutionalize bias through AI (Sabol et al., 2004). Working towards the Quintuple Aim of healthcare, ethical and equitable AI requires that we deliver on equity, transparency, and accountability from design to implementation.

Fairness and bias in AI

There is growing demand for AI programs that reflect the societal and cultural norms of the context in which they will be used. Fairness must not be viewed as a technical term that can be solved by simply including more data but must instead be deliberately addressed by understanding the explicit and implicit biases within AI systems and the equitability of the intended targets and actions. Even a rich, global data set such as Wikipedia has been shown to harbor gender stereotypes (Wagner et al., 2015). The inherent cultural, social, and educational biases that we hold are reflected in the way data is collected, processed, and organized, and algorithms are created. Gender, sexual orientation, racial, ethnicity, and age biases in both datasets and algorithmic predictions have been well described (Cirillo et al., 2020; Diaz et al., 2018; Wang et al., 2020). In healthcare, the downstream effects of biased and unfair AI systems automate inequity, structural determinants of health, and healthcare disparities.

How bias occurs in data sets

The adage "garbage in, garbage out" in AI is used to describe the need for valid and reliable datasets. Because the outputs from an AI program are extrapolated to large populations, the data collected and used to train algorithms must be as representative and inclusive of the population it intends to treat. This begins with sourcing geo-diverse data, as geographic homogeneity in data source harbors local social and cultural values that may not reflect that of a greater nation or state (Kaushal et al., 2020). Secondly, under-representation of gender, class, sex, and age will further reduce generalizability. Consider one example of using AI to diagnose skin cancers in the United States (Esteva et al., 2017). Under-representation of African Americans in the training image database could lead to inaccurate program outputs, false diagnoses, and ineffective treatments for a disenfranchised population (Zou & Schiebinger, 2018). For cases of rare diagnoses or conditions that effect the very elderly population, unbalanced data sets may be all that are available; therefore, an effort to slow down and struggle with inclusivity is crucial.

Additionally, humans' choices of how to label and organize data is non-random and vulnerable to personal opinions and assumptions (Sun et al., 2020). When the amount of well-annotated data is limited, model reliability may be compromised (Wang et al., 2020).The use of binary classifications—categories such as race, sexual orientation, and gender—is especially prone to bias and potential high error rates in algorithm output (Ngan & Grother, 2015; Vyas et al., 2020). In 2020, Google removed gender labels from its image recognition AI after realizing the impossibility of inferring someone's gender by appearance alone and in recognition that doing so perpetuates bias. However, deletion of labels may simultaneously undermine accuracy. Thus, efforts must be taken to mitigate the bias that is bound to exist within programs.

How bias occurs in algorithm processing

One of the challenges to any AI algorithm is that the program translates aspects of human life, a history of discrimination, and disparate healthcare into quantifiable inputs and outputs. This favors phenomena, inputs, and outputs that are more easily measured but that may not be more important.

One of the most notable demonstrations of how algorithmic bias can propagate healthcare disparities comes from Obermeyer and colleagues. They examined a large healthcare insurance database from over 49,000 Americans from which a commercial risk-prediction tool is derived, with outputs provided to large healthcare systems and payers. The outputs may be applied to 200 million insured Americans. The intent of the algorithm was to identify the "highest risk" patients—individuals with the most complex healthcare needs—who would benefit from additional resources and care coordination, without stratifying for race. Many healthcare systems use this and similar programs to allocate finite resources, reduce costs, and improve healthcare outcomes. Due to compounding socioeconomic factors, African American patients have reduced access to healthcare in the United States and appeared to have reduced healthcare spending; the algorithm under-identified the percentage of African American patients who would benefit from additional help. When the data was disaggregated, African American patients were spending much more on emergency room visits and dialysis compared to elective procedures than their Caucasian counterparts. The

authors then reformulated the algorithm to avoid using cost as a proxy for need and were able to eliminate racial bias in the algorithm (Obermeyer et al., 2019).

In the outpatient world, machine learning algorithms are employed to enhance clinic efficiency and minimize patient no-show rates. "Predictive overbooking" is one common framework to minimize clinician overtime and idle time while also minimizing patient wait times. Investigation of this algorithmic model revealed that a patient's no-show rate correlated with African American race. The scheduling system consequently resulted in African American patients facing 30 percent longer wait times. The authors then developed an alternate "race-aware objective" that instead of minimizing wait times of all patients, minimized wait times from the racial group expected to wait the longest, and were able to address both fairness and efficiency in this way (Samorani & Blount, 2020).

Solutions and practices to combat and ameliorate bias and lack of representativeness

A thoughtful and collaborative approach is needed to combat and ameliorate bias and lack of representativeness in clinical AI products. Equitable AI products are essential to promoting equitable health outcomes. This requires trust, transparency, and accountability.

Transparency

Transparency is a cornerstone of equity. The adoption of AI tools in an opaque fashion erodes patient trust or perpetuates harm. AI developers must be held responsible to disclose data collection methodologies, algorithmic processing, and decisional criteria to end-users. Providers need to know when to modify a model's recommendations based on each individual patient at a given point of care. While some AI methods are "black box" by design and outperform those in which the algorithm is interpretable, care must be taken to build other mechanisms, such as visualizations and parallel modeling from the same data, to provide sufficient transparency for the given use case.

For some applications, if providers cannot understand how or why a model arrived at a diagnosis or recommendation, patients and providers cannot be expected to trust AI-driven recommendations. Finally, providers should be transparent about the use of AI in clinical decision support. Communicating with patients about the use of AI technologies may increase patient trust and acceptance and improve shared decision-making.

Accountability

To mitigate potential "weapons of math destruction" (O'Neil, 2016), accountability is needed across the various steps of AI program development including, but not limited to, data sourcing and processing, algorithm assumptions and development, and clinical implementation. Regulatory oversight is needed to standardize and report high-quality data collection and ensure that the data is representative of the general population. Developers must acknowledge the source of their data, recognize that data labels may reflective structural inequalities, and consider relabeling of data to reduce bias (Mullainathan & Obermeyer,

2017). When possible, public datasets should be used to minimize privacy breaches. Ideally, datasets should be interconnected and accessible by stakeholders, ensuring appropriate safeguards.

Algorithms should report performance metrics across different subpopulation categories to avoid hidden AI bias, even when the model achieves similar accuracy across subgroups (Noseworthy et al., 2020). This enables developers to discover areas where their models may not be generalizable and in turn, perform additional model training on diverse demographic populations. Auditing for bias from public–private firm partnerships is needed to ensure safe, real-world applications.

Finally, in the vast majority of clinical settings, it is unlikely AI will become completely autonomous and will most likely augment existing healthcare practices. The level of autonomy desired and allowed will be predicated on the clinical risk calculus and scale of operation. At their core, algorithms are built from human experience and knowledge. Therefore, anti-bias and implicit bias training, and diverse representation in algorithm developers and providers is needed to ensure that equity and inclusion are prioritized.

Legal and Institutional Regulation of AI-Driven Medical Devices and Diagnostic Tools

National and international regulatory frameworks for healthcare AI

One of the primary concerns of AI regulation is ensuring patient safety and assuring that clinical AI tools are efficacious. Ensuring patient safety can be challenging because many AI tools are constantly learning, drawing upon complex datasets to generate conclusions that may be better refined with further data (He et al., 2019). Health AI may be regulated by agencies that also regulate medical devices, such as the United States' Food and Drug Administration (FDA), the European Medicines Agency (EMA), and other national agencies. While some other countries likewise regulate clinical AI products, such as China's National Medical Products Administration, most low- and middle-income countries recognize approval from the FDA or the European Union as a substitute for their own regulatory review. Therefore, this section will concentrate on European, American, and Chinese regulation of healthcare AI.

European Union regulations overview

The European Union is still in the process of articulating regulations for products in this area, focusing more on infrastructure than the regulation of specific products (Minssen et al., 2020). EMA and the Heads of Medicines Agencies (HMA), in collaboration on a Joint Big Data Taskforce, published two reports in 2019 and 2020 articulating priority actions for the European regulatory agencies in their approaches to data use. Similarly,

the European Commission published a 2020 White Paper on a European approach to AI, along with a report on the safety and liability of healthcare AI. In the White Paper, the European Commission emphasized that trustworthiness is a prerequisite for the successful implementation of AI (European Commission, 2020). The High-Level Expert Group on AI created by the European Commission further articulated that the essential regulatory requirements for AI implementation are "(1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability" (European Commission, 2019). On a more practical level, all medical devices, including those with embedded AI algorithms, must comply with the relevant EU regulatory frameworks to be marketed to European consumers, regardless of their residency (Minssen et al., 2020).

United States of America regulations overview

The regulation of healthcare AI is very much in flux in the United States. The FDA distinguishes between AI as general wellness products and AI as medical devices, choosing to regulate only the latter. The 21st Century Cures Act likewise excludes software intended for "maintaining or encouraging a healthy lifestyle" that is "unrelated to the diagnosis, cure, mitigation, prevention, or treatment of a disease or condition" (U.S. Congress, 2016). Only AI products that meet the definition of device under the Federal Food and Drug Cosmetic Act are regulated by the FDA, requiring pre- and post-marketing regulatory controls.

People's Republic of China regulations overview

China is heavily invested in developing AI, including healthcare AI, with its State Council's 2017 release of the "New Generation Artificial Intelligence Development Plan," a unified document that articulates China's policy objectives for AI (Roberts et al., 2021). Similarly, regulatory agencies, such as the Center for Medical Device Evaluation (CMDE) of the National Medical Products Administration (NMPA), have established AI working groups and published some preliminary regulatory guidance for AI developers and manufacturers (Li, 2020). Overall, China has embraced the potential of AI, including its ability to extend health care services to underserved communities, and therefore is supportive of bringing these products to market (Roberts et al., 2021).

International regulations

Another key player in this space is the International Medical Devices Regulators Forum (IMDRF), a voluntary global group of medical device regulators from jurisdictions such as Australia, Canada, Japan, and the United States. The IMDRF has continued the work of the Global Harmonization Task Force on medical devices, with an eye to encouraging the development and adoption of new technologies and encouraging regulatory convergence across jurisdictions. In 2019, IMDRF published a framework for regulating software as a medical device that is applicable to many healthcare AI products (IMDRF, 2014). Other international efforts include the OECD Principles on AI, released in 2019 and signed by 42 countries.

American and European approaches to key regulatory issues in healthcare AI

Expanding on the general national and international regulatory landscape of AI in healthcare applications, the following is a more detailed analysis of the regulatory strategies adopted by the United States of America and the European Union, two of the leaders in the development of clinical AI tools, in relation to three critical policy aspects: AI tool classification, adaptive algorithms, and data governance.

Classification by risk-based assessment

Both the FDA and the European approaches classify medical devices on the basis of risk to the user. Which category a clinical AI tool will fall into will depend on its uses and the framework used by the relevant agency. For example, most medical devices that included embedded AI that monitor physiological processes will be classed as Class IIa or higher (meaning medium to higher risk) in the EU regulatory framework (Minssen et al., 2020). Decision-support AI tools will likely be categorized depending on the seriousness of the health care situation involved rather than on the basis of functionality (IMDRF, 2014). The classification of a particular algorithm will determine the level of evaluation, scrutiny, and oversight the product will receive from the relevant regulators.

Regulatory considerations for online learning and adaptive algorithms

One unique challenge facing the regulation of healthcare AI devices and clinical AI tools is how to evaluate and oversee algorithms. Learning, or adaptive, algorithms continually update their predictions based on new data and inputs as opposed to locked algorithms, which are trained on a specific data set and then apply those predictions to future data without refining the predictions based on the newer data.

While the benefits of learning algorithms are clear, especially for patients who may not be well represented in training data sets, there are also difficulties in ensuring the quality of these algorithms. The updates that form the backbone of learning algorithms happen without human supervision and are almost continuous, making it hard to understand when the algorithm has materially changed from the version evaluated by regulators (Babic et al., 2019). Learning algorithms are also susceptible to data quality issues, such as physicians manipulating billing codes to ensure access to certain treatments and services for their patients (Finlayson et al., 2019). At present time, the comparative standard for algorithms remains that of an expert clinician and does not focus on what matters most in real-world settings: long-term outcomes of a significantly larger population. Regulating bodies and governmental policies should also consider associated public health risks of overdiagnosis and increasing false positives.

The European Union has struggled in its approach to regulating learning algorithms. In a February 2020 White Paper, the European Commission acknowledged that learning algorithms can pose some unique concerns, including cybersecurity issues (European Commission, 2020). Unfortunately, this White Paper does not provide solutions or suggest a regulatory approach to these products. Part of the challenge is that EMA provides a

regulatory framework laid on top of the systems of each individual country. Therefore, it can be more challenging to monitor post-approval data for AI products across different countries' data tracking systems (European Medicines Agency, 2018). This means that EMA likely needs to focus on evaluation of learning algorithms prior to approval, a challenge for products that are constantly in flux (Cohen et al., 2020).

Until April 2019, the FDA chose to regulate only locked algorithms, defined as "an algorithm that provides the same result each time the same input is applied to it and does not change with use" (IMDRF, 2013). The FDA then issued a discussion paper and request for feedback that proposed a regulatory approach to overseeing adaptive algorithms embedded in medical devices that it called "total produce lifecycle (TPLC) regulatory approach" (IMDRF, 2013). The TPLC approach includes a pre-certification program and requirements for manufacturers of learning algorithms to submit a "predetermined change control plan," or an articulation of anticipated modifications, and an "algorithm change protocol," or an articulation of the associated methodology used to implement continuous changes. Although this proposal has not yet been fully implemented, in early 2021 the FDA issued an action plan outlining its intent to carry out most of the recommendations in the initial discussion paper (U.S. Food & Drug Administration, 2021).

Data governance and regulation

Data privacy and security present significant challenges for AI regulation. Some of the specific concerns that AI regulators must address include (1) patient, family, and consumer engagement and education; (2) de-identification of patient data; (3) data sharing; and (4) ownership of data. Some of these challenges can be addressed through education, some through existing regulations, and some will require creative solutions from stakeholders. To illustrate the challenges of data governance for healthcare AI, this section will focus on concerns raised by use of personal data and the regulatory solutions posed by European and American regulators.

Individuals are often sensitive to the use of their personal data, both for training and other purposes. A common solution is to limit the use of individually identifiable data to specific applications, either defined by law or granted by individual authorization or consent. Of course, that can create a significant impediment to sharing data, especially in the context of research and development of new technologies. Individuals are also often concerned with whether their data is being shared and who "owns" their data. This is no exception when it comes to the underlying health data that powers healthcare AI. In response to the challenges of obtaining authorizations for data use, most data regulation regimes allow for much freer use of anonymized or de-identified data sets, which aggregate data from multiple individuals while removing key data that could serve to identify the people within the data set. These data sets are the ones most commonly used to train and create AI, especially in the medical context.

The General Data Protection Regulation (GDPR) governs the use of data in the European Economic Area. The GDPR is a general data protection statute and does not limit itself to health applications but is certainly applicable to healthcare AI. In fact, the GDPR classifies health-related data as "sensitive data," with heightened protections (Forcier et al., 2019). The GDPR has several levels of de-identification, distinguishing between pseudonymization and true anonymization. The level of de-identification used determines which GDPR regulations are imposed. Recital 26 of the GDPR defines anonymized data as "data

rendered anonymous in such a way that the data subject is not or no longer identifiable." Full anonymization means that the data is no longer considered personal and, therefore, is no longer regulated by the GDPR. This means that developers and researchers would have much more leeway working with anonymized data. Truly anonymized data, however, is difficult to achieve (European Commission, 2016). Pseudonymized data, as defined in Article 4(5) of the GDPR as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information," may be a more realistic goal. Article 6(4)(e) of the GDPR allows for pseudonymized data to be used for reasons beyond the purpose for which the data was originally collected. Parties working with such pseudonymized data should be careful that this data would not be "reasonably likely" to result in re-identification.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) governs the use of data included in the medical records of physicians, clinics, and hospitals but not data that is generated in less conventional settings, such as some digital health applications (Cohen & Mello, 2018). Therefore, while virtually all clinical AI tools will fall under HIPAA because they will be used by traditional health care providers, some health-related AI products will fall outside of the scope of HIPAA. This may mean that there are few to no specific consumer data protections for American users of these products (Cohen & Mello, 2018), unless there are state-specific statutes such as the California Consumer Privacy Act (Forcier et al., 2019).

HIPAA requires organizations seeking to use patient data to obtain individual authorizations before disclosing or using their health information, which may be onerous for developers of AI who customarily use large datasets to train their algorithms. A significant exception to the authorization requirement, however, is the use of de-identified data sets. These data sets have key elements of information removed, such as names and social security numbers, with the goal of making it impossible to identify individuals within the larger data set. Unlike the GDPR, HIPAA does not have levels of anonymization. Data is either de-identified or it is not.

The use of de-identified data to protect patient privacy raises some challenges for AI developers and users. First, the great the level of de-identification, generally the greater the loss of data utility and value (Hintze, 2017). De-identified data sets may not allow AI tools to reach their maximum utility and development. Additionally, it is unclear whether de-identification prevents individuals from being re-identified and therefore protects patient privacy (Benitez & Malin, 2010). For example, in 1996, researchers were able to use publicly available de-identified hospital records along with identified voter registration records to re-identify the medical records of the then-governor of Massachusetts (Benitez & Malin, 2010). In the age of data aggregation, re-identification is looming larger and larger as a threat to patient privacy. In the United States, courts have begun to examine data aggregation as a threat to patient privacy protections such as in the class action lawsuit *Dinerstein v. Google* (Cohen & Mello, 2019).

Recommendations and Hypotheses for a Successful Future for AI in Healthcare

As evidenced by the remarkable strides in AI-related healthcare applications and increasing sophistication in AI-health infrastructures, the next era of patient care and healthcare

innovation will inevitably feature AI as a central tenet. Equitable and effective implementation of emerging AI technologies in healthcare in the long-term and across disparate global settings, however, faces significant logistical, legal, and ethical barriers related to governance. Here, we conclude our arguments with three broad recommendations/directives, specifically in relation to national and international governance, for the promotion of a successful future for AI in healthcare.

1) *Addressing bias in and promoting equity through AI in healthcare*. AI algorithms in all fields, including healthcare, learn from data in an opaque process. Thus, race, gender, or other historical and sociodemographic biases in healthcare delivery can be instantiated in and perpetuated by such algorithms—without the knowledge or intent of developers or practitioners. This is further exacerbated when a diverse set of stakeholders from patients to practitioners, representing the diversity of the populations AI plans to serve are not at decision making tables when defining and delivering AI solutions. Moreover, access to high quality care, including state-of-the-art AI technologies, is often limited to the those who have the privilege of race, gender, cisgender, wealth, ableism, etc. Therefore, it is crucial to prioritize equity and inclusion when defining and delivering governance related to:
   (a) The future of AI in healthcare rests on the continual evolution of AI and its integration into health systems, while designing and evaluating interventions based on the Quintuple Aim (National Academy of Medicine, 2019). This includes patient outcomes, cost reduction, population impact, clinician wellness, and above all equity and inclusion.
   (b) Engaging a diverse and empowered set of stakeholders to inform AI development and testing, including robust incorporation of broader sociodemographic factors beyond race as model covariates.
   (c) Initiatives in medical education to create a more diverse technical and medical workforce, equipped with quantitative and procedural knowledge to effectively and equitably employ AI in their practices.
2) *Prioritizing proactivity and flexibility in algorithm development*. Not only do medical workflows and healthcare management procedures differ significantly by nation, local and global standards of care are continuously evolving in step with scientific progress and new research discoveries. In turn, regulatory oversight of healthcare-related AI should emphasize active maintenance and responsive updating of algorithms, to ensure that the large investments being made in this field today produce durable and far-reaching innovations. This could be balanced in gray areas or areas that are changing too quickly or need flexibility with accreditation and certification processes through large non-profit organizations with appropriate stakeholders to promote trust, safety, and equity. Such initiatives may ultimately mandate some degree of "explainability" and transparency to facilitate more effective and targeted algorithm modifications.
3) *Ensuring harmonization of regulatory frameworks and application domains*. AI-powered healthcare tools are generally developed in limited settings—designed for a very specific given task and trained with data from one hospital or a hospital system, for instance. Furthermore, that algorithm is produced within the confines of a given legal and national regulatory framework. A challenging and critical facet of

AI-health governance involves guidance on how to approach (a) safe inter-specialty and inter-hospital use of certain AI tools; and (b) international implementation of algorithms, specifically in relation to harmonization of differing legal requirements on the national and international levels. A closely related issue is that of AI algorithm interaction. As AI technologies scale and become increasingly prevalent in healthcare workflows, predictive patterns may interact and conflict, resulting in algorithm degradation.

References

Abdullah, R., & Fakieh, B. (2020). Health care employees' perceptions of the use of artificial intelligence applications: Survey study. *Journal of Medical Internet Research 22* (5), e17620. https://doi.org/10.2196/17620.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research 22* (6), e15154. https://doi.org/10.2196/15154.

Babic, B., Cohen, I. G., Evgeniou, T., & Gerke, S. (2021, January–February). When machine learning goes off the rails. Harvard Business Review. https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2019). Algorithms on regulatory lockdown in medicine. *Science 366* (6470), 1202–1204. https://doi.org/10.1126/science.aay9547.

Baharuden, A., Isaac, O., & Ameen, A. (2019). Factors influencing big data & analytics (BD&A) learning intentions with transformational leadership as moderator variable: Malaysian SME perspective. *International Journal of Management and Human Science 3* (1), 10–20. https://ejournal.lucp.net/index.php/ijmhs/article/view/791.

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association 17* (2), 169–177. https://doi.org/10.1136/jamia.2009.000026.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research 81*, 134–148. http://proceedings.mlr.press/v81/chouldechova18a.html.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digital Medicine 3* (1), 81. https://doi.org/10.1038/s41746-020-0288-5.

Cohen, I. G., Evgeniou, T., Gerke, S., & Minssen, T. (2020). The European artificial intelligence strategy: Implications and challenges for digital health. *The Lancet Digital Health 2* (7), e376–e379. https://doi.org/10.1016/s2589-7500(20)30112-6.

Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. *JAMA 320* (3), 231. https://doi.org/10.1001/jama.2018.5630.

Cohen, I. G., & Mello, M. M. (2019). Big data, big tech, and protecting patient privacy. *JAMA 322* (12), 1141–1142. https://doi.org/10.1001/jama.2019.11365.

Correia, R., Kon, F., & Kon, R. (2008). Borboleta: A mobile telehealth system for primary homecare. *SAC '08: Proceedings of the 2008 ACM Symposium on Applied Computing*, 1343–1347. https://doi.org/10.1145/1363686.1363998.

Crowley, R. J., Tan, Y. J., & Ioannidis, J. P. A. (2020). Empirical assessment of bias in machine learning diagnostic test accuracy studies. *Journal of the American Medical Informatics Association 27* (7), 1092–1101. https://doi.org/10.1093/jamia/ocaa075.

D'Agostino, S. R. B., Grundy, S., Sullivan, L. M., Wilson, P., & Group, for the C. R. P. (2001). Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. *JAMA 286* (2), 180–187. https://doi.org/10.1001/jama.286.2.180.

Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G., & Matheny, M. E. (2019). Comparison of prediction model performance updating protocols: Using a data-driven testing procedure to guide updating. *AMIA Annual Symposium Proceedings 2019*, 1002–1010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153129/.

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., Mandryk, R., Hancock, M., Perry, M., Cox, A., & Gergle, D. (2018). Addressing age-related bias in sentiment analysis. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173986.

Dopp, A. R., Parisi, K. E., Munson, S. A., & Lyon, A. R. (2020). Aligning implementation and user-centered design strategies to enhance the impact of health services: Results from a concept mapping study. *Implementation Science Communications 1* (1), 17. https://doi.org/10.1186/s43058-020-00020-w.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542* (7639), 115–118. https://doi.org/10.1038/nature21056.

European Commission. (2016). *Article 29 working party—European Commission*. https://ec.europa.eu/newsroom/article29/news-overview.cfm.

European Commission. (2019). *Ethics guidelines for trustworthy AI—Shaping Europe's digital future*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

European Commission. (2020). White Paper on artificial intelligence European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Medicines Agency. (2018). *A common data model for Europe?—Why? Which? How?* https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf.

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2020). Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Annals *of Operations Research 294* (1–2), 567–592. https://doi.org/10.1007/s10479-018-2818-y.

Fenech, M. E., & Buston, O. (2020). AI in cardiac imaging: A UK-based perspective on addressing the ethical, social, and political challenges. *Frontiers in Cardiovascular Medicine 7*, 54. https://doi.org/10.3389/fcvm.2020.00054.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science 363* (6433), 1287–1289. https://doi.org/10.1126/science.aaw4399.

Forcier, M. B., Gallois, H., Mullan, S., & Joly, Y. (2019). Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers? *Journal of Law and the Biosciences 6* (1), 317–335. https://doi.org/10.1093/jlb/lsz013.

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J. C., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., Graaf, J. de,

Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., Kleijn, D. P. V. de, Lonn, E. M., . . . Ruijter, H. M. den. (2015). Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLOS ONE 10* (7), e0132321. https://doi.org/10.1371/journal.pone.0132321.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine 25* (1), 30–36. https://doi.org/10.1038/s41591-018-0307-0.

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., & Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association 27* (12), ocaa088. https://doi.org/10.1093/jamia/ocaa088.

Hintze, M. (2017). Viewing the GDPR through a de-identification lens: A tool for compliance, clarification, and consistency. *International Data Privacy Law 8* (1), 86–101. https://doi.org/10.1093/idpl/ipx020.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., & Palade, V. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence 49* (7), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5.

Hunt, R., & McKelvey, F. (2019). Algorithmic regulation in media and cultural policy: A framework to evaluate barriers to accountability. *Journal of Information Policy 9*, 307–335. https://doi.org/10.5325/jinfopoli.9.2019.0307.

IMDRF. (2013). Software as a medical device (SaMD): Key definitions. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf.

IMDRF. (2014). "Software as a medical device": Possible framework for risk categorization and corresponding considerations. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

Johnson, K. B., Wei, W., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and Translational Science 14* (1), 86–93. https://doi.org/10.1111/cts.12884.

Kakadiaris, I. A., Vrigkas, M., Yen, A. A., Kuznetsova, T., Budoff, M., & Naghavi, M. (2018). Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of the American Heart Association 7* (22), e009476. https://doi.org/10.1161/jaha.118.009476.

Kaushal, A., Altman, R., & Langlotz, C. (2020). Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA 324* (12), 1212–1213. https://doi.org/10.1001/jama.2020.12067.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine 17* (1), 195. https://doi.org/10.1186/s12916-019-1426-2.

Koshy, A. N., Sajeev, J. K., Nerlekar, N., Brown, A. J., Rajakariar, K., Zureik, M., Wong, M. C., Roberts, L., Street, M., Cooke, J., & Teh, A. W. (2018). Smart watches for heart rate assessment in atrial arrhythmias. *International Journal of Cardiology 266*, 124–127. https://doi.org/10.1016/j.ijcard.2018.02.073.

Laï, M.-C., Brian, M., & Mamzer, M.-F. (2020). Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *Journal of Translational Medicine 18* (1), 14. https://doi.org/10.1186/s12967-019-02204-y.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science 343* (6176), 1203–1205. https://doi.org/10.1126/science.1248506.

Lee, D. J., Macke, S., Xin, D., Lee, A., Huang, S., & Parameswaran, A. G. (2019). A human-in-the-loop perspective on AutoML: Milestones and the road ahead. *IEEE Data Engineering Bulletin*, *42*, 59–70. https://www.semanticscholar.org/paper/A-Human-in-the-loop-Perspective-on-AutoML%3A-and-the-Lee-Macke/24bb20608df98963d9c706c3e44db0c126573264.

Li, Jacky. (2020). AI regulatory developments for medical devices in China. *Cisema*. https://www.cisema.com/en/ai-regulatory-developments-and-planning-issued-by-china-regulator/.

Maron, R. C., Weichenthal, M., Utikal, J. S., Hekler, A., Berking, C., Hauschild, A., Enk, A. H., Haferkamp, S., Klode, J., Schadendorf, D., Jansen, P., Holland-Letz, T., Schilling, B., Kalle, C. von Fröhling, S., Gaiser, M. R., Hartmann, D., Gesierich, A., Kähler, K. C., . . . Thiem, A. (2019). Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer 119*, 57–65. https://doi.org/10.1016/j.ejca.2019.06.013.

McCradden, M. D., Sarker, T., & Paprica, P. A. (2020). Conditionally positive: A qualitative study of public perceptions about using health data for artificial intelligence research. *BMJ Open 10* (10), e039798. https://doi.org/10.1136/bmjopen-2020-039798.

McGrail, K. M., Ahuja, M. A., & Leaver, C. A. (2017). Virtual visits and patient-centered care: Results of a patient survey and observational study. *Journal of Medical Internet Research 19* (5), e177. https://doi.org/10.2196/jmir.7374.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature 577* (7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Minssen, T., Gerke, S., Aboy, M., Price, N., & Cohen, G. (2020). Regulatory responses to medical machine learning. *Journal of Law and the Biosciences 7* (1). https://doi.org/10.1093/jlb/lsaa002

Mudgal, K. S., & Das, N. (2020). The ethical adoption of artificial intelligence in radiology. *British Journal of Radiology 2* (1), 20190020. https://doi.org/10.1259/bjro.20190020.

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review 107* (5), 476–480. https://doi.org/10.1257/aer.p20171084.

National Academy of Medicine. (2019). *Artificial intelligence special publication*. https://nam.edu/artificial-intelligence-special-publication/.

Ngan, M. & Grother, P (2015). Face recognition vendor test (FRVT) performance of automated gender classification algorithms. US Department of Commerce, National Institute of Standards and Technology. *NIST Interagency/Internal Report (NISTIR)*. https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., & Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nature Medicine 26* (9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y.

Noseworthy, P. A., Attia, Z. I., Brewer, L. C., Hayes, S. N., Yao, X., Kapa, S., Friedman, P. A., & Lopez-Jimenez, F. (2020). Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis. *Circulation: Arrhythmia and Electrophysiology 13* (3). https://doi.org/10.1161/circep.119.007988.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366* (6464), 447–453. https://doi.org/10.1126/science.aax2342.

O'Neil, C. (2016). *Weapons of math destruction*. Penguin Random House.

Paranjape, K., Schinkel, M., Panday, R. N., Car, J., & Nanayakkara, P. (2019). Introducing artificial intelligence training in medical education. *JMIR Medical Education 5* (2), e16048. https://doi.org/10.2196/16048.

Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental 2* (1), 35. https://doi.org/10.1186/s41747-018-0061-6.

Petitgand, C., Motulsky, A., Denis, J.-L., & Régis, C. (2020). Investigating the barriers to physician adoption of an artificial intelligence-based decision support system in emergency care: An interpretative qualitative study. *Studies in Health Technology and Informatics 270*, 1001–1005. https://doi.org/10.3233/shti200312.

Quinn, C. C., Clough, S. S., Minor, J. M., Lender, D., Okafor, M. C., & Gruber-Baldini, A. (2008). WellDoc™ mobile diabetes management randomized controlled trial: Change in clinical and behavioral outcomes and patient and physician satisfaction. *Diabetes Technology & Therapeutics 10* (3), 160–168. https://doi.org/10.1089/dia.2008.0283.

Rastogi, N., Gloria, M. J. K., & Hendler, H. (2015). Security and privacy of performing data analytics in the cloud: a three-way handshake of technology, policy, and management. *Journal of Information Policy*, *5*, 129. https://doi.org/10.5325/jinfopoli.5.2015.0129.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2019). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, *27*(3), 491–497. https://doi.org/10.1093/jamia/ocz192.

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J., on behalf of the SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI Extension. *BMJ 370*, m3210. https://doi.org/10.1136/bmj.m3210.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & SOCIETY*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2.

Roski, J., Gillingham, B. L., Just, E., Barr, S., Sohn, E., & Sakarcan, K. (2018, September 18). Implementing and scaling artificial intelligence solutions: Considerations for policy makers and decision makers. *Health Affairs*. https://www.healthaffairs.org/do/10.1377/hblog20180917.283077/full/.

Roski, J., Maier, E. J., Vigilante, K., Kane, E. A., & Matheny, M. E. (2021). Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocab065.

Sabol, W., Coulton, C., & Polousky, E. (2004). Measuring child maltreatment risk in communities: a life table approach. *Child Abuse & Neglect 28* (9), 967–983. https://doi.org/10.1016/j.chiabu.2004.03.014.

Samorani, M., & Blount, L. G. (2020). Machine learning and medical appointment scheduling: Creating and perpetuating inequalities in access to health care. *American Journal of Public Health 110* (4), 440–441. https://doi.org/10.2105/ajph.2020.305570.

Sarwar, S., Dent, A., Faust, K., Richer, M., Djuric, U., Ommeren, R. V., & Diamandis, P. (2019). Physician perspectives on integration of artificial intelligence into diagnostic pathology. *Npj Digital Medicine 2* (1), 28. https://doi.org/10.1038/s41746-019-0106-0.

Seneviratne, M. G., Hersch, F., & Peiris, D. P. (2018). HealthNavigator: A mobile application for chronic disease screening and linkage to services at an urban Primary Health Network. *Australian Journal of Primary Health 24* (2), 116. https://doi.org/10.1071/py17070.

Shi, S., He, D., Li, L., Kumar, N., Khan, M. K., & Choo, K.-K. R. (2020). Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. *Computers & Security 97*, 101966. https://doi.org/10.1016/j.cose.2020.101966.

Sohn, E., Roski, J., Escaravage, S., & Maloy, K. (2017, May 9). Four lessons in the adoption of machine learning in health care. *Health Affairs*. https://www.healthaffairs.org/do/10.1377/hblog20170509.059985/full/.

Solomon, D. H., & Rudin, R. S. (2020). Digital health technologies: Opportunities and challenges in rheumatology. *Nature Reviews Rheumatology 16* (9), 525–535. https://doi.org/10.1038/s41584-020-0461-x.

Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P. C., & Moors, E. H. M. (2020). Implementation of artificial intelligence (AI) applications in radiology: Hindering and facilitating factors. *European Radiology 30* (10), 5525–5532. https://doi.org/10.1007/s00330-020-06946-y.

Subbaswamy, A., & Saria, S. (2019). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics 21* (2), 345–352. https://doi.org/10.1093/biostatistics/kxz041.

Sun, W., Nasraoui, O., & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE 15* (8), e0235502. https://doi.org/10.1371/journal.pone.0235502.

Topol, E. (2019). *Deep medicine*. Basic Books.

Tsao, C. W., & Vasan, R. S. (2015). Cohort profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology 44* (6), 1800–1813. https://doi.org/10.1093/ije/dyv337.

Turakhia, M. P., Desai, M., Hedlin, H., Rajmane, A., Talati, N., Ferris, T., Desai, S., Nag, D., Patel, M., Kowey, P., Rumsfeld, J. S., Russo, A. M., Hills, M. T., Granger, C. B., Mahaffey, K. W., & Perez, M. V. (2019). Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple heart study. *American Heart Journal 207*, 66–75. https://doi.org/10.1016/j.ahj.2018.09.002.

Tzachor, A., Whittlestone, J., Sundaram, L., & ÓhÉigeartaigh, S. (2020). Artificial intelligence in a crisis needs ethics with urgency. *Nature Machine Intelligence 2* (7), 365–366. https://doi.org/10.1038/s42256-020-0195-0.

U.S. Congress. (2016). *H.R.34—114th Congress (2015–2016): 21st Century Cures Act*. https://www.congress.gov/bill/114th-congress/house-bill/34.

U.S. Food & Drug Administration. (2021, January). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. https://www.fda.gov/media/145022/download.

Vriesema, J. L. J., van der Poel, H. G., Debruyne, F. M. J., Schalken, J. A., Kok, L. P., & Boon, M. E. (2000). Neural network-based digitized cell image diagnosis of bladder wash cytology. *Diagnostic Cytopathology 23* (3), 171–179. https://doi.org/10.1002/1097-0339(200009)23:3<171::aid-dc6>3.0.co;2-f.

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine 383* (9), 874–882. https://doi.org/10.1056/nejmms2004740.

Wadhwa, V., Alagappan, M., Gonzalez, A., Gupta, K., Brown, J., Cohen, J., Sawhney, M., Pleskow, D., & Berzin, T. (2020). Physician sentiment toward artificial intelligence (AI) in colonoscopic practice: A survey of US gastroenterologists. *Endoscopy International Open 8* (10), e1379–e1384. https://doi.org/10.1055/a-1223-1926.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *ArXiv*. https://arxiv.org/abs/1501.06307.

Wall, J., & Krummel, T. (2019). The digital surgeon: How big data, automation, and artificial intelligence will change surgical practice. *Journal of Pediatric Surgery 55*, 47–50. https://doi.org/10.1016/j.jpedsurg.2019.09.008.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples. *ACM Computing Surveys (CSUR) 53* (3), 1–34. https://doi.org/10.1145/3386252.

Yin, K., Pandian, V., Kadimisetty, K., Zhang, X., Ruiz, C., Cooper, K., & Liu, C. (2020). Real-time colorimetric quantitative molecular detection of infectious diseases on smartphone-based diagnostic platform. *Scientific Reports 10* (1), 9009. https://doi.org/10.1038/s41598-020-65899-w.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature 559* (7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8.

CHAPTER 42

AI, FINTECH, AND THE EVOLVING REGULATION OF CONSUMER FINANCIAL PRIVACY

NIKITA AGGARWAL

Introduction

Over the last 50 years, consumer financial markets have undergone a dramatic transformation, from being predominantly face-to-face, paper-based, and bank-dominated, to becoming much more digital, automated, and decentralized. This phenomenon is frequently referred to as "fintech."[1] The rise of fintech was spurred by advances in computing in the late 1960s. We might refer to this period as "fintech 1.0." Fintech has expanded over the last 20 years, with the wider diffusion of the consumer Internet beginning in the 1990s, and the growth in mobile computing and social media since the early 2000s. Today, in the era of fintech *2.0*, consumers can open bank accounts, take out loans, invest and make payments, all through an app on their mobile phone. Increasingly, these apps are powered by "artificial intelligence" (AI)—in particular, algorithmic models built using machine learning (ML) and other statistical techniques, and large volumes of personal data. Consumer financial services are also increasingly likely to be offered by non-bank, non-financial institutions (Cornelli et al., 2020)—whether a peer-to-peer (p2p) lending platform like Prosper,[2] a mobile payment services provider like Venmo,[3] or the financial services arm of a Big Tech company like Apple.[4] As such, the boundaries of the financial sector have become much more blurred and diffuse.

As with all advances in information technology, fintech—and AI/data-driven finance specifically—is a double-edged sword. This narrative is well-known. On the one hand, fintech stands to benefit consumers and the economy, inter alia through lower costs, more personalized, and more convenient access to finance. On the other hand, fintech can harm consumers, inter alia by increasing the scope for illegal discrimination and by enabling firms to more effectively exploit vulnerable consumers using data-driven insights (Aggarwal,

2021b). Focusing on the growing use of AI and personal data in consumer finance, this chapter examines fintech through the less-analyzed lens of consumer financial privacy. Adopting a historical perspective, the chapter begins by examining the regulation of consumer financial privacy as it has evolved over the course of the last century, from the duty of bank confidentiality to cross-sectoral ("omnibus") data protection regulation. As will be demonstrated, this evolution has been shaped directly by the rise of fintech—specifically, the growth in the use of personal data for consumer financial decision-making. Finally, the chapter describes the recent rise of AI-driven finance, examines the implications of this trend for the future of consumer financial privacy, and lays out an agenda for further research on the regulation of consumer financial privacy.

To limit the scope of inquiry, the chapter focuses on the regulation of fintech and consumer financial privacy under English law. However, this analysis will also be relevant to other jurisdictions that are grappling with the challenge of regulating financial markets, and consumer financial privacy, in the era of fintech and AI. Furthermore, given the aims of this *Handbook*, the chapter focuses on the implications of AI, specifically, for consumer financial privacy. As such, it does not examine other important fintech developments that are relevant to consumer financial privacy: in particular, the rise of digital money and payments, including cryptocurrencies (Committee on Payments and Market Infrastructures, 2015; Nakomoto, 2008). Although not addressed by this chapter, further research in these areas will be essential.

Bank Confidentiality and the Origins of Consumer Financial Privacy

The origins of consumer financial privacy under English law lie in the common law duty of bank confidentiality. In its 1924 decision in *Tournier v National Union and Provincial Bank of England*,[5] the Court of Appeal held that banks have an implied contractual duty to keep confidential and not misuse the personal information of their customers. As a special instance of the general law of confidence, this "duty of bank confidentiality"—and the customer's corresponding right to confidence—extend to any information relating to the customer that has been generated within the bank–customer relationship, including both financial as well as non-financial information,[6] provided that it is non-trivial, has a confidential quality of which the bank has notice, and is not public or common knowledge.[7] The duty has since been extended to non-bank financial institutions, such as credit unions[8]—and could conceivably extend to other non-bank financial institutions (Cranston et al., 2018, p. 258).

There are various rationales for recognizing a duty of bank confidentiality as a legal and not simply moral duty (Stokes, 2011). The court in *Tournier* sought to strengthen protection for the confidentiality of the bank–customer relationship, and particularly the professional duty of confidentiality owed by banks to their customers (Aplin et al., 2012, pp. 200, 381ff). The court also observed that the duty of bank confidentiality was necessary to protect the reputation of a bank's customers and in turn their ability to access credit.[9] More fundamentally, the duty of bank confidentiality has been justified as necessary to protect the privacy

and autonomy of bank customers (Aplin et al., 2012, p. 200). Personal financial information can reveal many intimate details about a person (Levitin, 2018, p. 767ff). The duty has also been justified in the interests of maintaining public trust in banks (Cranston et al., 2018, p. 255)—and for preserving "the kind of society in which we want to live" (Jack, 1989, p. 34).

Nevertheless, it has long been recognized that there are good reasons for banks to disclose their customers' information in certain circumstances. Thus, the court in *Tournier* established four qualifications to the duty of bank confidentiality. These are: "(a) where disclosure is under compulsion of law; (b) where there is a duty to the public to disclose; (c) where the interests of the bank require disclosure; (d) where the disclosure is made by the express or implied consent of the customer."[10] Inter alia, these qualifications permitted banks to share customer information amongst themselves for credit referencing purposes. Of course, banks could also agree express confidentiality clauses with their customers, which is common in financial contracts (Cranston et al., 2018, p. 257).

Socio-technical developments since *Tournier* have led to an expansion of these qualifications, further narrowing the scope of the duty of bank confidentiality. In particular, rapid advances in computing from the late 1960s onwards—which significantly reduced the cost of collecting, storing and sharing information—facilitated the development of large consumer (credit) databases and more sophisticated statistical and computerized credit scoring systems. These technological advances enabled the growth of the consumer credit market, particularly the unsecured credit market (Crowther, 1971, p. 366ff; Thomas et al., 2017, p. 4). In turn, as the consumer credit market grew and took on an increasingly important role in the economy, greater value came to be placed on sharing customers' credit information between banks for the purposes of assessing customers' credit risk (Younger, 1972, p. 78; Jack, 1989, p. 33). Likewise, the use of credit data to develop credit scoring models became a standard feature of credit risk management and consumer lending, increasingly justifying the encroachment on bank-customer confidentiality. The expansion of the consumer credit market also increased the value to firms of using customer data to analyze customers' preferences and targeting them with financial offers (Thomas et al., 2017, pp. 5, 9).

More broadly, there was a gradual shift in opinion regarding the value of bank–customer confidentiality relative to the public interest in sharing information, a shift that accelerated in the wake of 9/11 and the ensuing War on Terror (Godfrey et al., 2016, p. 174). Today, financial institutions are required by domestic and international laws to share customer information with government agencies in order to combat crimes, such as tax evasion (OECD, 2017) and money laundering and terrorism financing (AML/CFT),[11] and to facilitate financial supervision (Basel Committee on Banking Supervision, 1981, 1990).[12]

It should be noted, however, that information shared in these contexts—whether for credit referencing or criminal enforcement purposes—is generally still subject to a duty of confidentiality on the part of the recipient, and conditions limiting the use of that information (Cranston et al., 2018, pp. 262–263; OECD, 2012).[13] Thus, although greater information-sharing by financial institutions has diluted the duty of bank confidentiality, it has not completely eroded it. Rather, it has become more legitimate to use the information of financial customers in contexts outside of the bank–customer relationship.

Another significant development that has narrowed the duty of bank confidentiality is the changing structure of banking and financial markets: in particular, the growth of "universal banking" and the related shift from bank-based finance to non-bank and market-based finance. Universal banking—whereby financial groups offer a range of banking and

non-banking financial services, such as insurance—has led to more customer information being shared within financial groups, particularly to enable cross-selling of financial products and services to customers, as well as for credit referencing and risk management purposes (Goode, 1989, p. 270; Jack, 1989, p. 31). In tandem, there has been significant growth in non-bank, non-deposit taking financial intermediaries (often referred to as "shadow banks"), such as insurers, pension funds, mutual funds, mobile payment service providers, and more recently, p2p lending platforms (FSB, 2020). Technological advances in low-cost computing and high-speed communications were an important enabler of these developments (Armour et al., 2016, p. 436).

The rise of universal banking and non-bank finance fundamentally challenges the scope of application of, and justification for, the duty of bank confidentiality. Amongst other things, universal banking muddies the waters on what exactly constitutes the "bank–customer relationship," and information "arising from" that relationship, for the purposes of recognizing the duty of bank confidentiality. Furthermore, consumers do not typically enjoy the kind of close and continuing relationship with p2p lenders and other non-bank financial intermediaries that was characteristic of personal banking, and the banking profession, at the time of *Tournier* (Jack, 1989, para 2.18).

Data Protection Regulation and the Evolution of Consumer Financial Privacy

Paradoxically, the same forces that instigated the erosion of the common law duty of bank confidentiality over the course of the 20th century also spurred the development of information privacy law: under the rubric of "data protection regulation," and latterly, human rights law protecting the individual right to privacy.[14] This chapter focuses on the governance of consumer financial privacy under cross-sectoral data protection regulation. The UK first began to develop data protection regulation in the early 1970s. Since the mid-1990s, UK data protection regulation has been shaped heavily by developments in EU law, particularly the growth of EU data protection and fundamental rights law. These two phases in the evolution of data protection regulation, and thus the governance of consumer financial privacy in the UK, are examined in turn.

The early development of information privacy law in the UK

Although the UK had long been a signatory to various international treaties recognizing the right to privacy—notably, the Universal Declaration of Human Rights and the European Convention on Human Rights—it wasn't until the early 1970s that the government began to think seriously about the need for domestic public regulation of privacy, specifically "information privacy" (Bennett, 2008, pp.4, 6–9; Jentzsch, 2003, p. 8). Until that point, information privacy had been a primarily "private not public concern" (Norman, 1999), regulated through private contract law, and in consumer financial markets specifically, the common law duty of bank confidentiality—as previously discussed.

However, rapid advances in computing and information processing beginning in the late 1960s, particularly in consumer credit markets, increased concern about the potential misuse of personal data and harm to individuals.[15] It was against this backdrop that the government in the early 1970s established a series of parliamentary committees to examine privacy under English law. The most prominent of these were the Younger Committee on Privacy, which issued its recommendations in 1972 (Younger, 1972; the "Younger Report"), and the Lindop Committee on Data Protection, which issued its recommendations in 1978 (Lindop, 1978; the "Lindop Report").[16]

Importantly for understanding the evolution of consumer financial privacy under English law (and the regulation of information privacy more broadly), the Younger and Lindop inquiries marked a shift in focus from private law rights and duties to protect the confidentiality of non-public information shared in the context of a trusting or professional relationship—such as the bank-customer relationship—to a much broader set of rights, duties and safeguards for the processing of personal information, both public as well as non-public (Younger, 1972, para 38; Lindop, 1978, paras 2.02–2.04). With the Lindop Report, the framing of information privacy began to shift even further away from the concepts of "confidentiality" and "privacy," towards the concept of "data protection" that was more familiar to Continental Europe (UK Home Office, 1975, para 36; Lindop, 1978, para 2.02; Jentzsch, 2003, p. 8).

At the same time, the Younger and Lindop Committees were keen to emphasize the *costs* of regulating information privacy: in particular, cutting off the economic and social gains due to information processing—such as access to credit—and stymying promising innovation and economic growth (Younger, 1972 paras 27 and 583; Lindop, 1978, para 2.08). As such, neither committee, nor the Government at the time, favored the introduction of a broad-based statutory right to privacy under English law, nor detailed regulation of computer systems (UK Home Office, 1975, para 28). Instead, their main recommendations were the establishment of a "Data Protection Authority" and a set of "principles for handling personal information" by computers (Younger, 1972, paras 591–600; Lindop, 1978, para 21.09). This included principles of "data minimisation," "purpose limitation," and "data accuracy," as well as limited rights for data subjects to be told about information held about them.

In addition, the Younger and Lindop committees made several specific recommendations for safeguarding privacy in banking and consumer credit markets, including a new right for individuals to access on request, and object to, the information held about them by credit reference agencies (CRAs) (Younger, 1972, paras 301–310). This right was also recommended by the Crowther Committee and enshrined in the UK Consumer Credit Act 1974.[17] The committees furthermore emphasized the unmet potential of the law on breach of confidence for addressing privacy concerns (Younger, 1972, para 630; Jack, 1989).

It was not until the mid-1980s that the principles of information management and the licensing regime, first articulated in the Younger Report, would be enforced by statute pursuant to the UK Data Protection Act 1984 (DPA 1984). Ultimately, the UK government was compelled to enact data protection legislation more out of commercial and economic interest, rather than concern to protect the privacy of personal data, or to mitigate harms to consumers due to the misuse of personal data (Bennett, 2008, pp. 141–142). More particularly, it sought to avoid obstacles to the cross-border flow of data under the Council of Europe's (CoE) Convention on Data Protection.[18] As a result, the DPA 1984 closely tracks the CoE Convention (Warren and Dearnley, 2005).

Overall, the DPA 1984—and the Younger and Lindop Reports that preceded it—sought to *enable* the processing of personal data (specifically, "automatic"; i.e., computer processing), whilst putting in place limited guardrails to prevent harm arising from the misuse of personal data. While they strengthened consumers' rights in the use of their data, these rights—and the obligations imposed on data processors—were constrained by the overriding goal of capturing efficiencies from the processing of personal data. As such, the fundamental, deontological right to control the use of one's personal information—the right to "informational self-determination" in the German idiom—was not a foundational goal of the DPA 1984 regime (Chalton, 1997, p. 32), although it was foundational to many European data protection laws, notably German law (Lynskey, 2015, pp. 94–95).[19]

The DPA 1984 has since been superseded, first by the 1998 Data Protection Act (DPA 1998), which implemented the 1995 EU Data Protection Directive (DPD 1995) in English law,[20] and most recently, the EU General Data Protection Regulation (GDPR) and UK Data Protection Act 2018 (DPA 2018).*

The GDPR and Data Protection Act 2018

The GDPR and DPA 2018 regime establishes duties for firms (as "data controllers" and "data processors"), rights for consumers (as "data subjects"), and regulatory enforcement powers with respect to the processing of personal data. As with the DPA 1998 regime, processing of personal data is prohibited unless permitted under one of the "lawful" grounds for processing—for example, that the consumer has given their "explicit" consent to processing ("opt-in" rather than "opt-out"),[21] or the processing is necessary for the controller's "legitimate interests."[22] Several of these grounds mirror the qualifications to the common law duty of bank confidentiality. However, as observed earlier, the data protection regime is much broader than the common law of (bank) confidence, not least because it applies to *all* "personal data"—public as well as non-public, financial as well as non-financial—provided that the data relates to an identified or identifiable natural person.[23]

The GDPR/DPA 2018 regime introduces various additional duties for data processors and controllers. Inter alia, this includes the duty to carry out a "data protection impact assessment" (DPIA) where the intended data processing is likely to result in a "high risk to the rights and freedoms of natural persons,"[24] and to implement "data protection by design and default" based on an assessment of the "risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing."[25] The data protection principles have been enhanced with explicit principles of "integrity and confidentiality," "transparency," and "accountability," in addition to pre-existing principles such as "data minimization" and "purpose limitation."[26] More types of data are now classified as sensitive or "special category" (notably, biometric and genetic data), and thus subject to stricter controls on processing.

* Note that following Brexit, the GDPR (and related EU jurisprudence) have become part of English law as 'direct EU legislation' (European Union (Withdrawal) Act 2018, s 3). Accordingly, reference will be made to the GDPR/DPA 2018 regime, which represents the current data protection regime in the UK, although it should be noted that these regimes could diverge in the future. See e.g. UK Department for Digital, Culture, Media & Sport (2021).

In turn, consumers have various rights to access, and control the processing of, their data. Inter alia, they are entitled to access and correct errors in their personal data[27] (rights that are reinforced by UK consumer credit law, as discussed earlier).[28] Of particular relevance to AI-driven finance is the qualified right for consumers to object to a decision taken solely on the basis of "automated processing," and to obtain human intervention in the decision[29]—as well as the more general right to receive information about the existence and "logic involved" in automated individual decision-making.[30]

Consumers also have new and/or strengthened rights to have their personal data erased (the so-called "right to be forgotten"),[31] to be informed without delay of a data breach,[32] to be given information about data held about them (via a "subject access request" or "SAR"), and to receive their personal data in a "structured, commonly used and machine readable format" and "transmit those data to another controller" (the right to "data portability").[33] The parallel development of "Open Banking" under EU and UK payments regulation gives consumers a sector-specific right to data portability limited to their financial account data (Borgogno & Poncibò, 2019).[34] Of course, these rights are not unqualified. The right to be forgotten is subject to the data no longer being necessary for the purposes for which it was collected.[35] Likewise, there is a presumption that any subject access request for data from a CRA is limited to personal data relevant to the consumer's "financial standing," unless the consumer specifically requests access to additional data.[36]

From a normative perspective, the GDPR/DPA 2018 regime—as with the earlier DPA 1984 and DPD 1995/DPA 1998 regimes—is characterized by a duality between, on the one hand, the goal of enabling personal data processing and the benefits that arise from such processing; and, on the other hand, the goal of preventing harm to data subjects arising from the misuse of personal data (Lynskey, 2015, p. 76ff). However, the GDPR/DPA 2018 regime places greater emphasis than previous regimes on mitigating the harms of data processing. Notably, it emphasizes the protection of individual fundamental rights, in particular the fundamental right to the protection of personal data under EU law (the "right to data protection").[37] As Lynskey argues, the right to data protection is rooted in the right to information privacy i.e. it grants individuals *enhanced control* over more types of personal data, in more contexts (2015, p. 90).[38] Indeed, privacy scholars (at least in the West) have long considered individual control over personal data to be a key facet of information privacy (Westin, 1967; Allen, 2000).

As such, the GDPR/DPA 2018 regime appears to place greater importance on the *intrinsic* value of individual data control; i.e., the ability of data subjects (consumers) to control their personal data, as an end in itself. However, it should be acknowledged that the normative significance of individual data control under this regime remains contested.[†] In its intrinsic role, individual control over personal data embodies the spirit of the right to "informational self-determination," which as discussed earlier is rooted in the Kantian ideals of personal autonomy and dignity. On this view, information privacy-as-data-control is per se a necessary (but not sufficient) condition for consumers to construct their own identities, self-present, self-realize, and exercise their autonomy (Gavison, 1980; Rouvroy & Poullet, 2009; Floridi, 2011; Bernal, 2014).

[†] See, e.g., the recent decision of the UK Supreme Court in *Lloyd v Google* ([2021] UKSC 50), and related commentary.

Importantly, the norm of individual data control in this intrinsic, deontological sense does not prohibit the alienation or commodification of personal data (Lynskey, 2015, p. 238ff). Nor is it a norm of non-interference (cf. Warren & Brandeis, 1890) or secrecy of non-public information, as under the common laws of privacy and confidentiality (cf. Posner, 1981). It is less a form of protection for a pre-cultural, essential self (a negative liberty conception; Berlin, 1969), as it is constitutive of the self—marking the boundaries within which individuals and communities can choose and construct their own identities, develop their capacities for self-determination and pursue the conditions for human flourishing (a positive liberty conception; Cohen, 2013; Aggarwal, 2021a).

Privacy-as-individual-data-control also performs an *instrumental* role, which can be traced back to the DPA 1984 and 1998 regimes, examined earlier.[39] Instrumentally, individual control over the processing of personal data is a mechanism for mitigating the consequential harms (enabling consequential benefits) due to data processing. This includes tangible harms involving financial loss, physical or psychological injury[40]—for example, identity theft or data profiling leading to unfair discrimination.[41] It also includes intangible privacy harms (Lynskey, 2015, pp. 210–227)—for example, the chilling effects and general feeling of powerlessness due to constant surveillance and behavioural profiling (Swire, 1999; Calo, 2011; Zuboff, 2019).

Of course, individual rights of control over data processing, and the corresponding obligations of data processors/controllers, are not the only mechanisms for limiting the harms (enabling the benefits) due to data processing, and thus achieving the instrumental goals of information privacy. The GDPR/DPA 2018 data protection regime also seeks to mitigate these harms through duties imposed on data processors and controllers, such as the duty to implement "privacy by default and design," the data protection principles, and the stricter requirements for consumer consent to data processing.

As will be examined next, consumer finance is increasingly being driven by data and AI/ML. This presents fresh challenges, as well as opportunities, for the regulation of consumer financial privacy.

Fintech, AI and the Future of Consumer Financial Privacy

As noted at the outset of this chapter, advances in computing beginning in the late 1960s spurred the digitization and datafication of financial markets, or fintech 1.0. The wider diffusion of the consumer internet beginning in the late 1990s, and the growth of mobile computing and social media in the early 2000s heralded the current era in fintech, or fintech 2.0. The earliest innovations in fintech 2.0 were in electronic payments and mobile money, pioneered by platforms like PayPal and M-Pesa. The contraction in lending by mainstream bank lenders following the 2008 global financial crisis provided an impetus for new technology-led financial startups to emerge, notably p2p lending platforms (Arner et al., 2015; FSB, 2017).

The fintech paradigm now embraces a wide range of entities, products, business models, and regulatory arrangements (FSB, 2017). This includes Big Tech companies, such as Apple

and Alibaba/Ant Financial, that are rapidly expanding their consumer finance offerings (Frost et al., 2019; Zetsche et al., 2017). It also includes traditional banks and financial institutions that increasingly rely on digital technology, and work in partnership with fintech startups and Big Tech companies to deliver financial services (Aggarwal, 2021b).

A key trend under fintech 2.0, which is highly salient to the evolution and regulation of consumer financial privacy, is the growing use of data analytics—particularly using AI/ML methods and "alternative data"—to personalize and automate consumer financial services. As society has become increasingly networked, digitized, and datafied (Mayer-Schönberger & Cukier, 2013), the volume of data that consumers generate has grown exponentially. There have also been notable improvements in data storage capacity, computational power, and processing speed—particularly due to the repurposing of "graphic processing units" for ML models (Marcus & Davies 2019, p. 43), and the development of more powerful and efficient ML-optimized chips, such as Google's "tensor processing unit."[42] Collectively, these technological developments have enabled much larger and more accurate ML models to be built—particularly "deep learning" (DL) models (LeCun et al., 2015). Whilst there have also been theoretical advances in ML methods, most of the fundamental methods were developed several decades ago—until recently, they were simply missing the data and hardware needed to power them (Goodfellow et al., 2015, pp. 12–22).

Data analytics, especially using AI/ML and alternative data, is being widely adopted in consumer finance—in everything from financial marketing, customer identity verification, and anti-money laundering/fraud control to financial advice, credit scoring and underwriting (Bank of England, 2019; Bholat et al., 2020). So-called "alternative data" includes both non-traditional *financial* data, such as consumers' Netflix and rental payments,[43] as well as *non-financial* data. The latter includes consumers' social media activity, mobile phone usage (Björkegren & Grissen, 2020), health fitness activity, retail and online browsing behaviour—even data on how consumers interact with a lender's website (Berg et al., 2019). Alternative data is typically less structured and more feature-rich ("high-dimensional") than traditional credit data.

What are the implications of increasingly AI and data-driven finance for consumer financial privacy, as regulated by the GDPR/DPA 2018? On the one hand, the use of AI/ML appears to undermine consumers' ability to control their personal data, thus diminishing their *intrinsic* financial privacy. Although both the GDPR/DPA 2018 and platforms such as Open Banking offer consumers greater control over whether, and if so which, third parties can process their personal data, there are technical, legal, and cognitive limits to consumers' ability to control the increasingly complex *inferences* and *predictions* that third parties can derive from that data, once they have been granted data processing authority. AI/ML methods, especially DL, can be used to parse large, unstructured, and high-dimensional datasets and identify features and patterns that are (more) relevant to predicting a target variable—such as a consumer's creditworthiness (Hurley & Adebayo 2017, p. 168ff). Importantly, these techniques can more accurately capture nonlinear relationships in data (Sadhwani et al., 2021).

Legally, consumers' data rights under the GDPR do not generally extend to the inferences and predictions derived from their data, inter alia using AI/ML methods (Wachter & Mittlestadt, 2019). More fundamentally, there are technical and cognitive limits to consumers' ability to control the inferences from their personal data. To begin with, there are well-known behavioural and cognitive weaknesses that undermine the effective exercise

by consumers of their data protection rights, and thus consumers' ability to "self-manage" their information privacy (Kahneman, 2011; Solove, 2013). In addition, the inferences underlying AI/ML models are mostly unobservable to consumers at the point of handing over their data (Cofone, 2020), and unintuitive—for example, the demonstrated correlation between the number of apps a consumer has installed on their mobile phone, and their perceived creditworthiness (Aggarwal et al., 2019). These insights are often obtained by repurposing and aggregating a consumer's data with other personal and non-personal data. Importantly, a person's data does not just reveal granular insights about them, but also about their friends, associates, and members of similar social/affinity groups (Mittelstadt, 2017; Madden et al., 2017). And once data is aggregated, commingled, and embedded in ML models, it is technically impossible to identify an individual consumer's personal data for the purposes of exercising their rights of control, such as the right to access, delete, or port their data (Öhman & Aggarwal, 2020, p. 8).

Moreover, many ML methods, particularly DL, are difficult to interpret, as a result of which even data processors are unable to explain the inferences underlying an AI/ML model. This problem could partly be overcome by mandating the use of more interpretable ("explainable") ML methods (Bracke et al., 2019; Chen et al., 2021), which could allow consumers to better understand the inferences and predictions derived from their data. However, interpretable ML will not in itself give consumers control over the inferences that might be derived from their data in the first instance —that is, what lenders and other financial firms are able to discover about them. As a result, consumers still lose some intrinsic control over their personal data, and thus intrinsic financial privacy, under the paradigm of AI-driven finance.

The implications of AI-driven finance for the *instrumental* goals of consumer financial privacy are more nuanced. On the one hand, AI-driven finance stands to support the instrumental goals of consumer financial privacy; that is, to protect consumers from harm and ultimately support their autonomy. From this perspective, even if (financial) consumers are less able to understand and/or control the inferences and predictions generated about them by an AI/ML model (i.e., they lose intrinsic financial privacy), they could nevertheless benefit from those inferences and predictions.

As noted, the use of alternative data and ML techniques for credit scoring can facilitate the more accurate assessment of consumer creditworthiness. This has been shown to reduce (direct) discrimination in consumer lending (Bartlett et al., 2019) and improve access to credit for certain consumers (Jagtiani & Lemieux, 2019). Arguably, improved access to affordable credit is autonomy-enhancing (Aggarwal, 2021a). These benefits are likely to be greatest for "thin file" or "credit invisible" consumers (Brevoort et al., 2015; Experian, 2021), who lack strong credit histories and for whom traditional credit scoring techniques are therefore less effective at predicting their creditworthiness. Importantly, the use of interpretable ML techniques could in certain contexts *limit* these benefits, by reducing the predictive accuracy of credit scoring models (Bono et al., 2021).

Yet, the use of AI in credit scoring, and consumer finance more broadly, could also harm consumers. Amongst other things, while some consumers stand to benefit from lower cost and better access to finance due to the use of AI/ML, other consumers could experience material harm due to costlier finance resulting from the improved observability of their characteristics and preferences (Fuster et al., 2020; Bartlett et al., 2019), which inter alia lower their perceived creditworthiness (Aggarwal, 2021a; Foohey & Greene, 2022) and/or

increase the scope for exploitation by firms based on their preferences and misperceptions, rather than "true" risk (Bar-Gill, 2019; Madden et al., 2017). This could perpetuate historic discrimination in credit markets against minority groups (Fuster et al., 2020).

The use of more interpretable AI/ML techniques could help to mitigate some of these data-driven harms by enabling consumers, firms and regulators to better understand the inferences underlying AI/ML models, and correct inaccurate and/or unfair decisions based on the model's predictions. However, interpretable AI/ML by itself will not mitigate concerns about harm due to *accurate* inferences: for example, where consumers face a higher cost of finance as a result of improved observability of their negative characteristics. Indeed, even consumers who supposedly gain in the immediate term—for example, through access to credit that is cheaper, in relative terms—could in fact be harmed in the long run due to credit that is still costly in absolute terms, particularly if the debt becomes unaffordable due to previously unforeseen financial shocks (a dilemma discussed extensively in the literature on high-cost credit and illustrated amongst other things by the collapse of the fintech lender Wonga in 2018).[44]

The extent to which these data-driven outcomes are considered unfair and harmful (or indeed, are even comparable and commensurable) raises complex moral questions that are beyond the scope of the present chapter (cf. Kiviat, 2019). More broadly, even if consumers are not materially harmed by AI and data-driven consumer finance, they arguably suffer non-material harm due to the feeling of constant surveillance and the chilling effect of knowing that "all data is credit data."[45]

Concluding Remarks

The concept and regulation of consumer financial privacy have evolved considerably over the course of the last century. As the common law duty of bank confidentiality, consumer financial privacy was originally concerned with protecting a consumer's confidence in their bank, and the relationship of trust between them—both to protect the relationship per se, and because financial information, being of an intimate nature that can reveal much about a person, could harm the consumer's credit and reputation if misused. It thus restricted banks and certain other financial institutions from sharing non-public information about consumers.

Although this duty was, from the beginning, subject to clear qualifications permitting banks to disclose customer information on public interest and business efficacy grounds, these qualifications have expanded over the course of the last century, resulting in a gradual erosion of the duty of bank confidentiality. To paraphrase Godfrey et al. (2016), the banker's duty of confidentiality is "a dying duty but not dead yet." The growth of the consumer credit and credit information markets in the post-war period were the first major inflection points in the evolution—and erosion—of the duty of bank confidentiality. The subsequent rise of universal and cross-border banking, as well as non-bank and market-based finance, were triggers for further erosion. Advances in digital technology were key enabling factors at each stage.

Yet, as bank confidentiality withered, privacy and data protection flourished. The growth of information privacy law since the 1970s, particularly under the rubric of data protection

regulation, significantly expanded the concept of consumer financial privacy under English law. As a result, the regulation of consumer financial privacy moved away from the paradigm of relational confidentiality—non-disclosure of private information arising from the bank-customer relationship—towards one of individual control and institutional safeguards over the processing of all personal information, regardless of context. Over time, this regime has come to place greater importance on the fundamental right of individual consumers to control the use of their personal data.

The rise of fintech 2.0, and, in particular, the growing use of AI for financial decision-making, presents both challenges as well as opportunities for consumer financial privacy. On the one hand, the ideal of consumer financial privacy as individual data control seems increasingly unworkable as more personal data is generated, collected, and processed for consumer financial decision-making. There are technical and cognitive limits to consumers' ability to control the inferences and predictions drawn from their data using AI/ML. As such, the growth of AI in consumer finance stands to compromise consumer financial privacy in the intrinsic sense.

On the other hand, the use of AI could support the underlying instrumental goals of consumer financial privacy. That is, despite losing control over their data, some consumers stand to benefit from the use of AI/ML techniques that, inter alia, mitigate discrimination and improve access to/lower the cost of finance. Indeed, to the extent that the use of less interpretable AI/ML methods offers more useful insights into a person's behaviour, and thus improves their access to finance, *less* control over personal data could be *more* beneficial to consumers in the instrumental, consequential sense.

At the same time, however, the use of AI in consumer finance can harm (other) consumers. Amongst other things, lenders can exploit consumers using more granular, data-driven insights about their preferences and misperceptions. In this regard, interpretable AI/ML could help to mitigate these data-driven harms by enabling consumers, firms and regulators to better understand the inferences underlying AI/ML models, and correct inaccurate and/or unfair decisions based on the model's predictions. Consumers also experience intangible harm due to the feeling of perpetual surveillance and behavioural profiling.

Clearly, consumer financial privacy is nuanced—the implications of AI-driven finance for consumer financial privacy even more so. Nevertheless, the existing UK/EU data protection regime appears increasingly inadequate for addressing the consumer privacy challenges of AI-driven finance, and AI-driven decision making more broadly (Zarsky, 2017). This chapter closes by identifying four normative questions that should be addressed by future research into, and regulation of, consumer financial privacy.[‡]

1) *First*, what relative value should be placed on intrinsic consumer (financial) privacy; that is, individual control over the use of personal data as an end in itself?
2) *Second*, are the (tangible) harms of data processing, such as the data-driven exploitation of vulnerable consumers, best mitigated by strengthening individual rights over personal data, and/or by strengthening the obligations of data processors, and the

[‡] Indeed, Brexit may offer a hidden opportunity for the UK to rethink data protection regulation for the era of AI. See, e.g., UK Department for Digital, Culture, Media & Sport (2021).

latter's enforcement thereof? Relatedly, what role can interpretable AI/ML play in mitigating the harms due to data processing?

3) *Third*, could "resurrecting" and strengthening the duty of bank confidentiality (and relatedly, the duties of care of financial institutions [FCA 2021]) provide a potential avenue for reform?
4) *Fourth*, should these questions be addressed under omnibus data protection regulation or sectoral, financial regulation, or both? In this regard, are existing provisions under sectoral regulation (such as consumer credit and payment services laws) that govern consumer financial privacy to be construed as *lex specialis*, and therefore given precedence over cross-sectoral data protection regulation in the case of conflict?

Notes

1. A portmanteau of "finance/financial services" and "technology". See http://www.americanbanker.com/bankthink/fintech-the-word-that-is-evolves-1077098-1.html. Of course, analog technology has been used in finance since prehistoric times, when written records—an early form of information technology—were used to evidence financial transactions, and abacuses were used for calculation (Arner et al., 2015).
2. https://www.prosper.com/about. However, note that at the time of writing many major p2p platforms are being acquired by, merged with or turned into banks, and ending their p2p lending/investing operations (Proud, 2021).
3. https://venmo.com/.
4. https://www.apple.com/apple-card/.
5. [1924] 1 KB 461.
6. n 5, as discussed by Aplin et al. (2012, p. 383).
7. *Attorney-General* v. *Guardian Newspapers Ltd. (No. 2)* [1990] 1 AC 109 (HL), 281–282 (per Goff LJ).
8. *Bodnar v Townsend* [2003] TASSC 148.
9. n 5 at 474 (per Bankes LJ).
10. n 5 at 473 (per Bankes LJ).
11. See also Article 7 of the 2001 Protocol to the EU Convention on Mutual Assistance in Criminal Matters [2001] OJ C326/1 (prohibiting EU Member States from invoking bank secrecy as a reason for refusing a request for mutual assistance from another Member State), as implemented in the UK by section 33 of the Crime (International Cooperation) Act 2003.
12. For example, section 175(5) of the UK Financial Services and Markets Act 2000 (FSMA) (setting out exceptions to the duty of (bank) confidentiality for disclosure of information to regulators or investigating authorities).
13. For example, FSMA ss 348–349 (restrictions on disclosure of confidential information by FCA, PRA, etc.).
14. Human Rights Act 2000, Article 8 (the right to private life). See generally Aplin et al. (2012, p. 200ff).
15. A similar reckoning was taking place at the same time in the EU. See European Commission, "Communication to the Council on a Community Data-Processing Policy" SEC (73) 4300 final, 13; European Parliament, "Resolution on the Protection of the Rights of the Individual

in the Face of Developing Technical Progress in the Field of Automatic Data Processing" [1975] OJ C60/48.

16. See also UK Home Office (1975).
17. ss. 157–159.
18. Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (CETS No. 108). See also UK Home Office (1982, para 25). Several other countries and international organizations also adopted data protection laws and principles during this period.
19. See the judgment of the German Constitutional Court of December 15, 1983, 1 BvR 209/83, BVerfG 65, 1 (the "Census decision").
20. European Parliament and Council Directive 95/46/EC of 24 October 1995 on the protection of individuals regarding the processing of personal data and on the free movement of such data [1995] OJ L281/23.
21. GDPR, Article 4(11). See also recital 32 GDPR (contrast with the thinner definition of consent under the DPD 1995, Article 2(h)).
22. GDPR, Articles 6 to 9.
23. GDPR, Article 4(1) (definition of personal data).
24. GDPR, Article 35.
25. GDPR, Articles 25 and 28(1) (indirectly extending the obligation to data processors).
26. GDPR, Article 5(1).
27. GDPR, Articles 14–16.
28. n 17.
29. GDPR, Articles 21 and 22.
30. GDPR, Articles 13(2)(f), 14(2)(g), 15(1)(h).
31. GDPR, Article 17. The right to data erasure under the previous data protection regime was much more limited (DPD 1995, Article 12(b); DPA 1998, s 14).
32. GDPR, Article 32.
33. GDPR, Article 20.
34. https://www.openbanking.org.uk/.
35. GDPR, Article 17.
36. DPA 2018, s 13(2).
37. GDPR, Recital 2; Article 16(1) TFEU ([2012] OJ C326/47 [Treaty of Lisbon Amending the Treaty on European Union and the Treaty establishing the European Community [2007] OJ C306/1]).
38. See references to data control in recitals 7, 68, 75, and 85 of the GDPR.
39. Building on Lynskey's (2015) distinction between "instrumental" and "conceptual" roles (pp. 177–228).
40. GDPR, recitals 75, 83, and 85 (referring to "physical, material and non-material harms") and Article 82(1) (referring to "material and non-material harms").
41. See GDPR recital 91, and Article 29 Data Protection Working Party 2016 (at 8–12 setting out the types of data processing, including credit scoring, that are likely to result in a high risk to fundamental rights and freedoms).
42. https://cloud.google.com/tpu.
43. https://www.experian.co.uk/consumer/experian-boost.html.
44. https://www.theguardian.com/business/2018/aug/30/wonga-collapses-into-administration.
45. Douglas Merrill, CEO of ZestAI, in Hardy (2012).

References

Aggarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2019). Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech. SSRN Scholarly Paper. http://dx.doi.org/10.2139/ssrn.3507827.

Aggarwal, N. (2019). Big data and the obsolescence of consumer credit reports. Oxford Law Faculty. https://www.law.ox.ac.uk/business-law-blog/blog/2019/07/big-data-and-obsolescence-consumer-credit-reports.

Aggarwal, N. (2021a). The norms of algorithmic credit scoring. *The Cambridge Law Journal 80* (1), 42–73. https://doi.org/10.1017/S0008197321000015.

Aggarwal, N. (2021b). Fintech credit and consumer financial protection. In Iris H-Y Chiu & Gudula Deipenbrock (Eds.), *The Routledge handbook of financial technology and law* . Routledge & CRC Press. https://www.routledge.com/Routledge-Handbook-of-Financial-Technology-and-Law/Chiu-Deipenbrock/p/book/9780367344146.

Allen, A. L. (2000). Privacy as Data Control: Conceptual, Practical and Moral Limits of the Paradigm. *Connecticut Law Review* 32, 861.

Aplin, T., Bently, L., Johnson, P., & Malynicz, S. (2012). *Gurry on breach of confidence: The protection of confidential information*. (2nd Edition). Oxford University Press.

Armour, J., Awrey, D., Davies, P., Enriques, L., Gordon, J. N., Mayer, C., & Payne, J. (2016). *Principles of financial regulation*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198786474.001.0001/acprof-9780198786474.

Arner, D. W., Barberis, J. N., & Buckley, R. P. (2015). *The evolution of fintech: A new post-crisis paradigm?* https://doi.org/10.2139/ssrn.2676553.

Article 29 Data Protection Working Party. (2016). Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in high risk" for the purposes of regulation 2016/679. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.

Bank of England and Financial Conduct Authority. (2019). Machine learning in UK financial services. https://www.fca.org.uk/publications/research/research-note-machine-learning-uk-financial-services.

Bar-Gill, O. (2019). Algorithmic price discrimination when demand is a function of both preferences and (mis)perceptions. *University of Chicago Law Review 86* (2) 217–254. https://chicagounbound.uchicago.edu/uclrev/vol86/iss2/12.

Bartlett, R. Morse, A., Stanton, R., & Wallace, N. (2019). Consumer lending discrimination in the Fintech era. NBER Working Paper No. 25943. https://ssrn.com/abstract=3063448.

Basel Committee on Banking Supervision. (1981). Banking secrecy and international cooperation in banking supervision. https://www.bis.org/publ/bcbs00f.htm.

Basel Committee on Banking Supervision. (1990). Information flows between banking supervisory authorities. https://www.bis.org/publ/bcbsc313.htm.

Bennett, C. J. (2008). *The privacy advocates: Resisting the spread of surveillance*. The MIT Press. https://www.jstor.org/stable/j.ctt5hhfb6.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2019). On the rise of fintechs—credit scoring using digital footprints. https://doi.org/10.2139/ssrn.3163781.

Berlin, I. (1969). Two Concepts of Liberty. In Isaiah Berlin, *Four Essays on Liberty* (pp. 118–172). Oxford University Press.

Bernal, P. (2014). *Internet privacy rights: Rights to protect autonomy*. Cambridge University Press.

Bholat, D., Gharbawi, M., & Thew, O. (2020). The impact of covid on machine learning and data science in UK banking. *Quarterly Bulletin*. https://www.bankofengland.co.uk/quarte

rly-bulletin/2020/2020-q4/the-impact-of-covid-on-machine-learning-and-data-science-in-uk-banking.

Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts loan repayment. *The World Bank Economic Review 34* (3), 618–634. https://doi.org/10.1093/wber/lhz006.

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy 37* (3), 585–617. https://doi.org/10.1093/oxrep/grab020.

Borgogno, O., and Poncibò, C. (2019). The day after tomorrow of banking: On fintech, data control and consumer empowerment. In Nikita Aggarwal et al. (Eds), *Autonomous systems and the law* (pp. 55–60). Beck.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. *Bank of England* Staff Working Paper No. 816. https://bit.ly/2TyIkod.

Brevoort, K. et al. (2015). Credit invisibles. *Consumer Financial Protection Bureau Office of Research Reports Series No. 15-1*. https://ssrn.com/abstract=3288821.

Calo, R. (2011). The boundaries of privacy harm. *Indiana Law Journal* 86(3) 1131–1162.

Chalton, S. (1997). The transposition into UK law of EU Directive 95/46/EC (the Data Protection Directive). *International Review of Law, Computers and Technology 11* (1), 25.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2021). A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. https://arxiv.org/abs/2106.02605.

Cofone, I. N. (2020). nothing to hide, but something to lose. *University of Toronto Law Journal 70* (1), 64.

Cohen, J.E. (2013) What is privacy for. *Harvard Law Review 126*, 1904.

Committee on Payments and Market Infrastructures. (2015). Digital currencies. https://www.bis.org/cpmi/publ/d137.htm.

Cornelli, G., Frost, J., Gambacorta, L., Rau, R., Wardrop, R., & Ziegler, T. (2020). Fintech and big tech credit: A new database. https://www.bis.org/publ/work887.htm.

Cranston, S. R., Avgouleas, E., Zwieten, K. van, Hare, C., & Sante, T. van. (2018). *Principles of banking law* (3rd Edition). Oxford University Press.

Crowther of Headingley, G. (1971). *Consumer credit. Report of the committee, chairman Lord Crowther. Vol 1 report*. Command Paper Cmnd. 4596.

Experian. (2021). Credit awareness week: millions still excluded from mainstream financial services despite drop in credit invisibles. https://www.experianplc.com/media/latest-news/2021/credit-awareness-week-millions-still-excluded-from-mainstream-financial-services-despite-drop-in-credit-invisibles/.

Financial Conduct Authority. (2021). A new consumer duty. Consultation Paper CP 21/36, December. https://www.fca.org.uk/publications/consultation-papers/cp21-36-new-consumer-duty-feedback-cp21-13-further-consultation.

Financial Stability Board (FSB). (2017). FinTech credit: Market structure, business models and financial stability implications. https://www.fsb.org/2017/05/fintech-credit-market-structure-business-models-and-financial-stability-implications/.

Financial Stability Board (FSB). (2020). Global monitoring report on non-bank financial intermediation 2020. https://www.fsb.org/2020/12/global-monitoring-report-on-non-bank-financial-intermediation-2020/.

Floridi, L. (2011). The informational nature of personal identity. *Minds and Machines 21* (4), 549.

Foohey, P., & Greene, S. S. (2022). Credit scoring duality. *Law and Contemporary Problems (forthcoming)*. https://ssrn.com/abstract=3992749.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., & Zbinden, P. (2019). BigTech and the changing structure of financial intermediation. https://www.bis.org/publ/work779.htm.

Fuster, A., Goldsmith-Pinkham, P. S., Ramadorai, T., & Walther, A. (2021). Predictably unequal? The effects of machine learning on credit markets. SSRN Scholarly Paper. https://ssrn.com/abstract=3072038.

Gavison, R. (1980). Privacy and the limits of law. *Yale Law Journal 89* (3), 421.

Godfrey, G., Newcomb, D., Burke, B., Chen, G., Schmidt, N., Stadler, E., Coucouni, D., Johnston, W., & Boss, W. H. (2016). Bank confidentiality—A dying duty but not dead yet. *Business Law International 17*, 173. https://heinonline.org/HOL/Page?handle=hein.journals/blawintnl17&id=181&div=&collection=.

Goode, R. M. (1989). The banker's duty of confidentiality. *Journal of Business Law*, May, 269–272.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Hardy, Q. (2012). Big data for the poor. *The New York Times*, July 5. https://bits.blogs.nytimes.com/2012/07/05/big-data-for-the-poor/.

Hurley, M., & Adebayo, J. (2017). Credit Scoring In The Era Of Big Data. *Yale Journal of Law and Technology 18* (1) 148–216. https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5.

Jack, R. B. (1989). Banking services: Law and practice. Report by the Review Committee. Command Papers Cm. 622.

Jagtiani, J., and Lemieux, C. (2019). Do fintech lenders penetrate areas that are underserved by traditional banks. https://ssrn.com/abstract=3178459.

Jentzsch, N. (2003). The regulation of financial privacy: The United States vs. Europe (No. 5). European Credit Research Institute. https://www.researchgate.net/publication/228538622_The_Regulation_of_Financial_Privacy_The_United_States_vs_Europe.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Kiviat, B. (2019). The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review 84* (6): 1134–1158. https://doi.org/10.1177/0003122419884917.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature 521*, 436 https://doi.org/10.1038/nature14539.

Levitin, A. (2018). *Consumer finance law: Markets and regulation*. Wolters Kluwer.

Lindop, N. (1978). Report of the committee on data protection. Chairman: Sir Norman Lindop. Presented to Parliament by the Secretary of State for the Home Department by command of Her Majesty. December. Command Paper Cmnd. 7341.

Lynskey, O. (2015). *The foundations of EU data protection law*. Oxford University Press.

Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, poverty, and big data: Matrix of vulnerabilities for poor Americans. *Washington University Law Review 95* (1), 53–126.

Marcus, G., & Davies, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Paragon.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. John Murray.

Mittelstadt, B. (2017). From individual to group privacy in big data analytics. *Philos. Technol 30*, 475–494. https://doi.org/10.1007/s13347-017-0253-7.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.

Norman, F. (1999). History of information policy. http://www.laits.utexas.edu/~anorman/long.extra/information/history.html.

Öhman, C., & Aggarwal, N. (2020). What if Facebook goes down? Ethical and legal considerations for the demise of big tech. *Internet Policy Review 9* (3) 1–21. https://policyreview.info/articles/analysis/what-if-facebook-goes-down-ethical-and-legal-considerations-demise-big-tech.

Organisation for Economic Co-operation and Development (OECD). (2012). Keeping it safe: The OECD guide on the protection of confidentiality of information exchanged for tax purposes. https://www.oecd.org/tax/exchange-of-tax-information/keeping-it-safe.htm.

Organisation for Economic Co-operation and Development (OECD) (2017). Standard for automatic exchange of financial account information in tax matters (2nd Edition). https://www.oecd-ilibrary.org/taxation/standard-for-automatic-exchange-of-financial-account-information-in-tax-matters-second-edition_9789264267992-en.

Posner, R. (1981). The economics of privacy. *American Economic Review 71* (2), 405.

Proud, L. (2021). Peer-to-peer lending's demise is cautionary tale. https://www.reuters.com/markets/asia/peer-to-peer-lendings-demise-is-cautionary-tale-2021-12-13/.

Rouvroy, A., & Poullet, Y. (2009). The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In Serge Gutwirth et al. (Eds.), *Reinventing data protection?* (pp. 45–76). Springer.

Sadhwani, A., Giesecke, K., Sirignano, J., (2021). Deep Learning for Mortgage Risk. *Journal of Financial Econometrics 19* (2), 313–368.

Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review 126*, 1880.

Stokes, R. (2011). The genesis of banking confidentiality. *The Journal of Legal History 32* (3), 279–294. https://doi.org/10.1080/01440365.2011.627153.

Swire, P. P. (1999). Financial privacy and the theory of high-tech government surveillance. *Washington University Law Review* 77(2), 461–512. https://openscholarship.wustl.edu/law_lawreview/vol77/iss2/6.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. (2nd Edition). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974560.

UK Department for Digital, Culture, Media & Sport. (2021). Data: A New Direction. https://www.gov.uk/government/consultations/data-a-new-direction.

UK Home Office. (1975). *Computers and privacy* (Cmnd. 6353).

UK Home Office. (1982). *Data protection, the government's proposals for legislation* (Cmnd 8539).

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review 2* 494–620.

Warren, A., & Dearnley, J. (2005). Data protection legislation in the United Kingdom: From development to statute 1969–84. *Information, Communication and Society* 8(2), 238–263. https://www.tandfonline.com/doi/abs/10.1080/13691180500146383.

Warren, S., and Brandeis, L. (1890). The right to privacy. *Harvard Law Review IV*, 193.

Westin, A. F. (1967) *Privacy and freedom*. Athenaeum.

Younger, K. G. (1972). *Report of the committee on privacy (Younger Committee)* (FV 82/6; Cmnd. 5012). The National Archives, Kew.

Zarsky, T. (2017). Incompatible: The GDPR in the age of big data. Seton Hall Law Review 47(2) 995–1020.

Zetzsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). From fintech to techfin: The regulatory challenges of data-driven finance. https://doi.org/10.2139/ssrn.2959925.

Zuboff, S. (2019). *The age of surveillance capitalism*. Profile Books.

SECTION IX

INTERNATIONAL POLITICS AND AI GOVERNANCE

JUSTIN B. BULLOCK

CHAPTER 43

DUELING PERSPECTIVES IN AI AND U.S.–CHINA RELATIONS

Technonationalism vs. Technoglobalism

JEFFREY DING

Introduction

In March 2018, the U.S. Trade Representative released its Section 301 report on China's unfair trade practices, marking one of the first volleys in a U.S.–China trade war centered on strategic technologies. As U.S.–China technological competition intensifies, mastery over AI has emerged as a focal point of contestation. Prominent U.S. policymakers are crafting national AI initiatives to secure American technological leadership. Likewise, sensitive to dependency on the U.S. for key components of AI systems, Chinese policymakers are planning to boost independent innovation in AI. As efforts to decouple the two technology ecosystems gain momentum, both governments seek to control AI as a national asset for wealth and power (Ding and Dafoe, 2021).

That same month, a group of Microsoft researchers published a landmark article demonstrating that machine translation had achieved human parity in Chinese-to-English news translation (Hassan et al., 2018). While the result marked an amazing advance in AI, the humans behind the process may be even more significant. Educated in countries ranging from Egypt to Poland, this group of 24 researchers collaborated across borders to produce this technological breakthrough, with half based at Microsoft's headquarters in Redmond, Washington, and the other half from Microsoft Research Asia (MSRA), a lab in Beijing, China. Fittingly, a translation tool that could open the door for more cross-border interactions was itself a product of global scientific and technological collaboration.[1]

These two publications, released a week apart, capture the dueling forces that shape the U.S.–China relationship in AI. The Section 301 report stands in for the *technonationalist* frame, which emphasizes interstate competition over technological assets. This analytic approach regards AI as a battleground for technological autonomy and dominance. The machine translation article, by contrast, represents the *technoglobalist* frame, which stresses

the globalization of innovation. This frame views AI as a source of a more interconnected U.S.–China relationship.

Of these two contending frames, technonationalism dominates analysis about the potential impact of AI on U.S.–China relations. Media, commentators, and policymakers frequently depict U.S.–China competition over AI as a Cold War arms race. One analyst stated that technonationalism in China "reached an apogee" with the release of its national action plan for the AI industry (Joseph, 2019, p. 200). This heightened sensitivity to the interaction between AI and interstate competition has unfolded alongside a broader resurgence of technonationalism in the vocabulary of policymakers and commentators. According to a LexisNexis search, total mentions of technonationalism in major news publications have spiked in recent years.[2]

In this chapter, I challenge the technonationalist perspective of U.S.–China competition over AI. First, I outline the various assumptions underlying technonationalism, mapping out how existing analyses of U.S.–China relations in AI are rooted in this frame. I then show how the globalization of innovation undermines the ability of states to engage in zero-sum competition over AI. In doing so, I examine cross-border flows in the essential drivers of AI development: data, computing capacity, research ideas, and talent. Lastly, I show that AI advances could enable more globalization, which only further dampens the technonationalist impulse.

AI and U.S.–China Competition Viewed Through a Technonationalist Lens

Before unpacking how technonationalism pervades narratives about U.S.–China relations in the AI era, it is important to clearly define the concept of technonationalism. Technonationalism is a tricky term to pin down (Green, 1995, p. 11). It is often selectively deployed as a prescriptive label to condemn states that actively seek to disrupt the existing technological hierarchy. Consider, for instance, a *Chatham House* expert commentary on the U.S.–China trade conflict. After describing the economic nationalist rhetoric and actions of both countries, the authors conclude, "We may be seeing a more *'technonationalist' China* as well as a *protectionist US*" (Summers & Kwok, 2018, emphasis added).[3] As David Edgerton, a prominent historian of science and technology notes, "The term techno-nationalism is used by Western analysts primarily in relation to Japan and now China, to describe a potentially, perhaps actually, dangerous thing" (Edgerton, 2011, p. 106).

For the purposes of this chapter, I take technonationalism as an analytic approach to comprehending the effects of technological change on international politics. Technonationalism centers the nation as the primary unit where technological innovations happen, where the consequences of technology materialize, and where drivers of technology are located. It also foregrounds technology as an asset for which states compete over to achieve national strength. Generally, scholars define technonationalist policies as those that promote indigenous innovation and protect domestic firms from foreign competition, aiming to contain scientific and technological advancements within the national context (Ostry & Nelson, 1995). Other scholars, however, have emphasized the flexibility of policies

that fit within the umbrella of "technonationalism."[4] These variants emphasize that states can adopt a mix of policy measures, including attracting foreign imports and investments in the short-term, so as to secure technological autonomy in the long-term.

Regarding how AI advances could influence U.S. and China relations, the vast majority of analysis adopts a technonationalist lens. Focused on which country will gain more from AI in terms of overall power, researchers often benchmark the national AI capabilities of the U.S. and China against each other (Ding, 2018a, 2019; Imbrie et al., 2020). The National Security Commission on Artificial Intelligence, a bipartisan body that makes recommendations for U.S. government AI policy, urges that the U.S. "must win the AI competition that is intensifying strategic competition with China" (National Security Commission on Artificial Intelligence, 2021, p. 2). Writing in the *China Quarterly of International Strategic Studies*, two Chinese international relations scholars regard AI as "becoming a main battlefield of competition between China and the United States," in which the technology's significance for national security makes such a contest "like a zero-sum game" (Wang & Chen, 2018, p. 257).

Most notably, commentators describe a burgeoning "AI arms race," linking current U.S.–China competition over AI to a Cold War environment in which the U.S. and Soviet Union were engaged in an existential battle to develop the most advanced technologies."[5] Media and scholarly accounts of an AI arms race between the U.S. and China have proliferated in the past couple of years. Before 2016, articles rarely mentioned the phrase and there were fewer than 300 hits for "AI arms race" on Google; by 2018, a search for the term generated more than 50,000 results (Zwetsloot et al., 2018).[6] "The primary manifestation of AI nationalism," Kerry Mackereth writes, "is the intensifying rhetoric of the AI 'arms race,' which pitches AI development as a zero-sum game where the victor will not only control the most advanced AI technology, but also enjoy economic, political, and military dominance over all other nations" (Mackereth, 2021).

The Globalization of Innovation in AI

What are alternative approaches to thinking about the effects of AI on U.S.–China relations? Even as states regard scientific and technological capabilities as increasingly central to national power, the process of scientific and technological advance is becoming increasingly globalized. The globalization of innovation represents a shift in the spatial scope of new knowledge production toward the interregional, intercontinental, or global scale.[7] This phenomenon can be further decomposed into three sub-processes in which the constraints of geography have receded: (1) "Talent circulation"—the globalization of high-skilled talent flows in science and technology (S&T) domains; (2) "Strategic alliances"—the globalization of corporate R&D activities; (3) "Research collaboration"—the globalization of scientific and technological knowledge (Table 43.1).

All three forms of globalized flows played a crucial role in Microsoft's key breakthrough in machine translation. The researchers themselves embody the circulation of global talent. MSRA in Beijing—the company's second largest base which led the research program—encapsulates global innovation networks. Lastly, the ultimate publication was connected with insights and ideas from the global AI research community.

Table 43.1 A Tripartite Decomposition of the Globalization of Innovation

Sub-process	Increasingly globalized activities in innovation	Primary agent of globalization
Talent circulation	High-skilled talent flows in S&T domains	People
Strategic alliances	Offshore R&D, standards alliances	Corporations, especially multinational ones
Research collaborations	Dissemination of scientific and technological knowledge	Scientific communities, particularly those based at universities

All three sub-processes also shape the broader U.S.–China relationship in AI. First, in terms of talent circulation, flows of highly skilled migrants across borders is an important thread in the AI technical community. Human talent may be the most fundamental driver of AI development. Companies devote billions toward recruiting talented AI researchers, and governments, most notably the Chinese government, have set up talent programs to attract AI researchers to work in their countries (Ding, 2018a). The American AI ecosystem also benefits greatly from global talent flows. One analysis by the Center for Security and Emerging Technology found that over half of AI professionals in the U.S. were born abroad. Moreover, in 2016, Chinese nationals accounted for more than 20 percent of U.S. PhD graduates in computer science (Zwetsloot et al., 2019).

Again, Microsoft's lab in Beijing—once dubbed by *MIT Technology Review* as "the world's hottest computer lab" (Huang, 2004)—encapsulates this trend. Along with researchers at the Paulson Institute, I analyzed global flows of MSRA fellowship awardees, which provide scholarships and internship opportunities to PhD students at Asian universities. Specifically, we analyzed the career trajectories of MSRA fellows who received PhD funding at universities based in China, Taiwan, and Hong Kong in 2009 and 2010. According to publicly available data on 33 of these MSRA fellows, nearly half worked overseas after graduation (MacroPolo, 2019). While MSRA fellows go on to build China's most promising top AI startups and conduct AI research at China's top technology companies and universities, many also take on leading roles at U.S. technology giants.

Second, U.S. and Chinese AI companies form strategic alliances that can reshape what constitutes national interests. Technical standards-setting forums are an important vehicle. Foreign companies play a role in shaping Chinese domestic standards. China's technical committee for cybersecurity standards has started to allow foreign companies to participate in standardization work, as they are allowed to vote at the working group level in developing standards even in a relatively sensitive area (Office of the U.S. Trade Representative, 2018). In the AI field, foreign companies with bases in China, such as Intel and Panasonic, participated in a large-scale effort by the China Electronic Standardization Institute to coordinate AI standards (Ding et al., 2018).[8]

While there have been high-profile cases of China pushing technonationalist standards, more often than not Chinese firms' participation in standards-setting involves strategic alliances with many international firms. In facial recognition-enabled identity authentication, for instance, many major Chinese banks have partnered with international tech giants like Google and Samsung in a standards alliance that competes against two other standards

alliances led by Chinese tech giants Tencent and Alibaba, respectively.[9] Indeed, Chinese firms may back standards set by foreign technology leaders rather than those pushed by the Chinese state and other Chinese firms, resulting in a politics of standards defined by "increasingly complex and cross-cutting cleavages" (Yao & Suttmeier, 2004). In one high-profile case that came to light in May 2018, the Chinese firm Lenovo voted for a 5G technology standard led by U.S. firm Qualcomm when an alternative standard by Chinese firm Huawei was also available (Hersey, 2018).

These strategic alliances also involve cross-licensing. Especially for industries relevant to AI, in which technologies are pieces of complex systems, patents on different parts of these systems are owned by different firms. Thus, cross-licensing of patents is a necessity. If patent rights are strictly enforced, a race to the bottom would occur and no firm could benefit from their inventions (Ostry & Nelson, 1995). The Google–Tencent licensing deal, announced in January 2018, illustrates the rationale behind these deals. Google gets to embed its AI-related software and products into IP and technologies into Chinese products as a way to access the Chinese market, while Tencent, which has an AI-related patent count of only about six percent of Google's, gives its nascent AI lab a boost using Google's technologies (Chong, 2018).

Lastly, cross-border collaborations on AI research represent a third form of globalized innovation. To substantiate these information flows, I collected data on high-impact AI publications from 2007 to 2017. These publications were in the top one percentile of most cited articles from all AI articles published that same year.[10] A scientometric analysis of this corpus reveals the country collaboration networks that undergird the AI research community. These country collaboration networks, which measure the linkages between the countries based on the institutional affiliation of authors who co-author papers together, flesh out the porousness of national containers when it comes to AI-related research outputs. Co-authorship could also serve as a proxy for the extent of social ties between researchers from different countries.

Figure 43.1 shows the country collaboration network for the AI research community in a Fruchterman–Reingold layout, an intuitive format in which nodes (countries) that share more connections are closer to each other on the graph. The resulting network reveals that the U.S. and China have the densest co-authorship ties of all country pairs. Along with England and Germany, the U.S. and China occupy central positions in the AI research community. The U.S. is the closest to the very center of the network and has close collaborations with all the countries. Compared to the U.S., China has closer connections with Australia and Singapore, while the U.S. is closer to European countries like Switzerland.

Co-authorship networks among researchers at the Association for the Advancement of Artificial Intelligence (AAAI) Conference, a leading AI conference, confirm the interconnectedness of the American and Chinese AI research ecosystems. In 2016, Japan's Science and Technology Foresight Centre, part of its National Institute of Science and Technology Policy, reviewed publication trends in AAAI from the previous five years. Among the sample of 11 leading countries in AI, collaborations between U.S.- and China-based researchers were the most common co-authorship relationship by far (Koshiba, 2016, p. 6). Of the co-authored papers involving Chinese researchers, about 55 percent involved collaborations with American peers. As for co-authored papers involving American researchers, about 37 percent involved collaborations with Chinese peers.

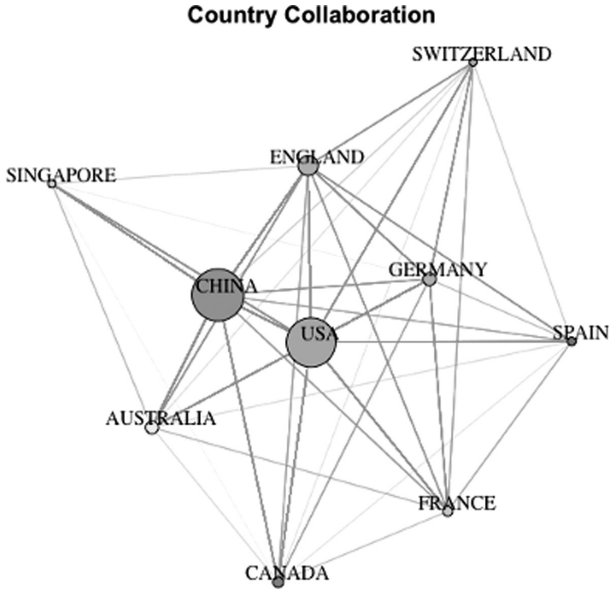

FIGURE 43.1 Country Collaboration

The Tension of Dueling Frames

The globalization of innovation is not slowing down. According to the most recent census statistics in 2010, there were about 28 million high-skilled migrants (those with a tertiary degree) residing in OECD countries, which represented an increase of nearly 130 percent since 1990 (Kerr et al., 2016). Crucially, firms are seeking access to innovative talent not only through attracting skilled migrants to their home bases but also by going to the homes of skilled migrants. This trend is reflected in the significant increases in co-inventions across all technology fields, multinational investments in foreign R&D centers, and R&D alliance activity (including joint ventures) (OECD, 2017). Finally, geographic constraints continue to recede in the world of scientific collaboration. Over the period from 2006 to 2016, the international co-authored proportion of all scientific and technical articles increased from 16.7 percent to 21.7 percent. In that same time frame, the share of academic publications co-authored by U.S. institutions and foreign institutions increased from 24.9 percent to 37.2 percent (National Science Board, 2018).[11]

The technoglobalist frame complicates efforts to treat U.S.–China competition in AI as a zero-sum game. Consider, for example, calls by prominent voices for U.S. companies to stop setting up AI labs in China because they help the Chinese military (Thiel, 2019). However, labs like MSRA, as key nodes in global innovation networks, benefit both the U.S. and China. On the one hand, MSRA has produced impressive results in applied AI fields such as computer vision, speech recognition, and translation. This directly benefits Microsoft, a key player in the U.S. tech industry. On the other hand, MSRA is also known as the Whampoa Military Academy of China's internet scene, a reference to the graduating school for many

Chinese military leaders who fought in 20th century conflicts. Indeed, the lab has trained over 4,800 Chinese interns, more than 100 professors teaching at top universities in China, and more than 500 professionals who are active in the Chinese IT industry, taking on executive and lead engineer positions in a variety of companies, including Baidu, Alibaba, Tencent, Huawei, Bytedance, and Lenovo (Wang, 2016).

There could be a rationale for severing such ties if it was clear that these labs markedly benefited one country more than another. Yet, it is difficult to calculate which country captures more relative gains from labs like MSRA. One comprehensive report on foreign-invested R&D centers in China, centered on information communications technology industries, concludes, "On balance, although foreign R&D centers are contributing to China's impressive recent high-tech growth and increasing competitiveness in ICT industries, they are contributing as much or more—under newly consolidated, wholly foreign-owned R&D enterprises—to foreign companies' high-tech development and production capabilities and, thus, to the US economy" (Walsh, 2003, p. xiv).

Do we see similar resilience to technonationalist impulses at work with respect to data flows, another crucial ingredient in AI systems? It is often said that, in the era of AI, data is the "new oil," and China is the "new Saudi Arabia" (Lee, 2018). Exerting national control over data flows, however, is much trickier than nationalizing oil supplies. In fact, U.S. AI companies and researchers regularly draw on Chinese datasets to build more precise AI systems. For instance, Datatang, a key data service provider in China which provides training datasets for AI companies, reported in 2017 that international companies accounted for half of its AI-related revenues (Ding, 2018b). In healthcare, U.S. researchers also collaborate with Chinese hospitals to use patient data for developing deep-learning-based diagnostic tools (Metz, 2019).

In recent years, the U.S. and China have both outlined policies based on the principles of data localization, which holds that data about a nation's citizens should be subject to reviews and restrictions before being transferred internationally. The U.S. government has attempted to block TikTok and WeChat, two Chinese-owned social media applications, due to concerns about their data collection practices (McKinnon & Leary, 2021). China has instituted data governance measures, including the 2017 Cybersecurity Law and a draft Personal Information Protection Law, that require handlers of critical and personal information to pass security assessments before transferring data abroad (Lee, 2021).

Regulators are aiming to ensure that companies protect sensitive data while still allowing them to derive the benefits of personalization afforded by AI algorithms—a tricky balance further complicated by the lack of technical expertise among those tasked with implementing data localization policies (Lee, 2021). After all, it is one thing to draft data localization initiatives, and another to fully enforce them.

Adding to the difficulty of enforcing data localization measures is the growing momentum toward data globalization. In 2016, a group of McKinsey Global Institute researchers computed trends in cross-border data flows over time. Using data sets from TeleGeography, a market research firm that maps usage of submarine optical fiber cables, they found that data flows have increased by a factor of 45 over the course of the past decade (Manyika et al., 2016). The United States is one of the central nodes of global digital networks, and China is becoming a more active participant in global data flows. Digital platforms facilitate almost one-fifth of Chinese imports and exports; a figure that was close to zero just two decades ago (Bughin & Lund, 2017).

Of all the inputs into AI systems, computing capacity may be the most controllable asset for states. The U.S. already implements export controls on various segments of the semiconductor supply chain, which are necessary for training and running AI algorithms (Fischer et al., 2021). Specifically, the U.S. has placed several Chinese AI companies, including Huawei, on an entity list that restricts exports of U.S.-origin semiconductor technologies.[12] These controls on critical technologies are effective because the U.S. can leverage central economic modes to restrict flows of semiconductors. Therefore, despite an increasingly global AI ecosystem, semiconductors have emerged as a possible "chokepoint" where states can "weaponize interdependence" (Farrell & Newman, 2019). Indeed, Chinese scholars also emphasize the need for major breakthroughs in core technologies such as chips so as to "prevent other countries from abusing weaponized networks to generate security threats" (Xu, 2021).

Still, even attempts to restrict semiconductor exports were partially thwarted by globally integrated networks. The initial controls did not cover chips designed by U.S. companies but manufactured in Taiwan or South Korea. When controls were expanded to semiconductor manufacturing equipment (SME), an industry more concentrated in the U.S., American SME companies complained that these actions would boost global competitors who could continue selling to the Chinese market (Brown, 2020a). After Huawei's addition to the entity list, companies like Google and Qualcomm pressured the U.S. government to provide exceptions and roll back restrictions (Malkin, 2020). Crucially, this push back differed from technological competition during periods where global integration was less robust. Chad Brown, a researcher at the Peterson Institute for International Economics, concludes, "Although the US semiconductor industry shared concerns about Chinese policies, it came out strongly against the US tariff actions. This response was distinct from earlier periods, when US trade restrictions on semiconductor imports from Japan, South Korea, and Taiwan emerged from direct requests by US firms" (Brown, 2020b, pp. 374–375).

Technoglobalism Taken Seriously

At times, the technonationalist frame misses the significant, independent effect of technology itself. Most scholarship on the intersection of AI and politics emphasizes how political circumstances, such as structural economic forces and interstate competition, will affect AI, while neglecting consideration of how AI itself will affect politics. While technoglobalism does not tell the whole story, throughout history, the impacts of revolutionary technologies have disrupted the conduct of relations between nations, often binding them more closely together in different ways.

The telegraph certainly revolutionized interactions between peoples, dropping communication times between Britain and India from six months in the 1830s to the same day in the 1870s (Buzan & Lawson, 2013). However, electronic telegraphy also presented significant coordination problems, such as one described by Ruggie concerning what happened to messages when they reached a border (Ruggie, 1993). Eventually, the European communications complex evolved from a series of bilateral treaties into several multilateral arrangements and, finally, into the International Telegraph Union (ITU) (Ruggie, 1993).

The ITU was the first standing intergovernmental organization, representing "the rise of permanent institutions of global governance" (Buzan & Lawson, 2013). This trend holds for other technologies as well; Murphy (1994) notes that a new generation of international organizations emerged as a regulatory response to revolutionary communication technologies, citing the ITU, the Radiotelegraph Union in 1906, and the International Telecommunications Satellite Organization in 1964. But the ITU set the precedent for the shape of multilateral arrangements to come: it codified rules of the road regarding the network of telegraph lines connecting Europe (and later the world), it established a permanent secretariat to administer those rules, and it convened periodic conferences to review the rules (Ruggie, 1993).

The telegraph's effects on global interdependence extended beyond measurable effects like the decreased delays in sending and receiving diplomatic messages. As Buzan and Lawson's (2013) text on the global transformation in the 19th century argues, the telegraph created a qualitatively different interaction capacity, linked peoples in a more intensely connected global economy, and generated a "19th century discourse" that shrank the world and allowed humanity to view themselves as one global body. These intangible effects of virtually instantaneous communication must be recognized.

In certain respects, innovations in AI also hold the potential to bind nations more closely together. One clear example is the domain of translation. Within just the past few years, AI techniques have resulted in significant leaps forward in machine translation. In 2016, Google Translate overhauled its phrase-based statistical machine translation system with one based on a neural network, resulting in "overnight improvements roughly equal to the total gains the old one had accrued over its entire lifetime" (a decade) (Lewis-Kraus, 2016). MSRA's breakthrough in automated Chinese–English translation occurred in 2018. Momentum is not slowing down any time soon. Per one analysis of a popular repository for computer science working papers, neural machine translation output hit record highs in April 2018 (Dino, 2018).

Because the nuances of language are ever-shifting and reliant on a broad and deep source of historical and cultural knowledge, machine translation may never be completely "solved." Still, significant improvements in machine translation are already making an impact on supply chains and productivity gains in the business world (Dino, 2018). Similar to other technologies, AI applications like machine translation could significantly slash the cost of transmitting information internationally by lowering transaction costs. For instance, researchers have argued that the rise of internet technologies helped propel the disaggregation of the state and the rise of transgovernmental networks as significant forces in international governance (Manulak & Snidal, 2020). The NSCAI report also notes that machine translation could "transform the way we communicate across geographic and cultural barriers, enabling business, diplomacy, and free exchange of ideas" (NSCAI, 2021).

Machine translation could shrink the world in a variety of ways. One simple application is improved understanding of Chinese trade barriers. The 2017 USTR report on foreign trade barriers contained a section on how China has not made available translations of all its trade-related laws, regulations, and other measures in one or more of the WTO languages (Office of the U.S. Trade Representative, 2017). This gap in translations of Chinese regulations could be largely mitigated by improvements in machine translation. Second, similar to internet technologies, machine translation could also lead to a marked reduction

in international communications, leading to a more integrated global system. Finally, similar to the intangible consequences of the telegraph outlined by Buzan and Lawson, improvements in machine translation could enable humanity to view themselves more as one global body, as people have more widespread access to popular works from other countries.

Despite the potentially transformative impact of machine translation, it is rarely mentioned in deliberations about the effect of AI on U.S.–China relations. One reason why, as this chapter outlines, is that the dominance of the technonationalist frame has unduly influenced discussions about U.S.–China relations in the era of AI. Moving beyond who wins and loses in AI opens up so many other interesting questions. For example, previous work on transnational communities of nuclear scientists has shown that they played an essential role in de-escalating the risks of nuclear war in the Cold War (Evangelista, 1999). A technoglobalist frame suggests that global networks of AI researchers could also prove pivotal in mitigating the risks of AI-enabled accidents and other risks. Taking technoglobalism seriously is essential to rebalancing discussions about how AI could impact the U.S.–China relationship.

Conclusion

No single analytic framework can capture all the nuances of how the U.S. and China will engage and compete in the AI domain. This chapter has identified a bias toward the technonationalist frame in existing scholarship on this topic. It has also demonstrated that the globalization of innovation is present in all the drivers of AI development, including data, talent, research, and computing capacity. This trend not only curtails the ability of states to control and protect AI assets but also generates uncertainty about whether actions to decouple the two ecosystems are net beneficial for either side. Lastly, I point out that continuing advances in AI could facilitate greater globalization and interdependence between the U.S. and China.

To be clear, I am not arguing that analysis of AI on U.S.–China relations should be centered on technoglobalism. Rather, a diversity of approaches is ideal. Future research should explore how states balance these two competing forces as they attempt to secure national interests in AI. Further analysis is also needed on how cross-national differences in the relationship between the state and domestic AI companies could dampen or intensify technonationalism. Lastly, the process of unpacking technoglobal and technational forces could also be expanded to countries other than the U.S. and China. Hopefully, this chapter provides a useful foundation and conceptual toolkit for such future research.

Acknowledgments

For helpful comments and input, I thank Miles Brundage, Justin Bullock, Allan Dafoe, Sophie-Charlotte Fisher, Carrick Flynn, Ben Garfinkel, Jade Leung, Cullen O'Keefe, Matthijs Maas, Brian Tse, Helen Toner, Waqar Zaidi, Baobao Zhang, Remco Zwetsloot, and an anonymous reviewer. This work was supported by the Berkeley Existential Risk Initiative.

Notes

1. For a similar story about the rollout of Skype Translator, see Kennedy (2018).
2. There were 46 mentions of the term "techno-nationalism" across major news publications in LexisNexis in 2018. From 2001 through 2003, the term received no mentions.
3. On the deep roots of Chinese technonationalism, see Feigenbaum (2003).
4. See Kennedy's "pragmatic techno-nationalism" (2013); Segal and Kang's "open techno-nationalism" (2006); the "technohybridism" described in Keller and Samuels (2003), or the "instrumental techno-nationalists" highlighted by Kennedy et al. (2008).
5. See, for example, Zhong and Mozur (2018).
6. According to reports of this impending U.S.–China AI arms race, the weapons China will use to win the AI arms race encompass supercomputers and quantum technologies (Barnes & Chin, 2018), which are unproven and have unclear linkages to AI development, an abundance of data useful for narrow applications of facial recognition and precision medicine (Lucas & Waters, 2018), which do not necessarily translate into useful military applications, and Chinese children (Beard, 2018), who are people not weapons.
7. This draws from the definition of globalization used in Avant et al. (2010).
8. Ding et al. (2018).
9. The standards alliances include Fast Identity Online Alliance (FIDO) Tencent's SOTER initiative.
10. This dataset and replication materials are available upon request. I used the bibliometric R-package to conduct this scientometric analysis.
11. One important caveat to all this is that the globalization of innovation is not truly global in scope but rather skewed toward connections among a group of rich, developed countries with the U.S. at the center and an emerging group of countries including China and India.
12. Some of these entities have been placed on the export blacklist for their involvement with human rights abuses in China's Xinjiang province.

References

Avant, Deborah D., Finnemore, Martha, & Sell, Susan K. (2010). *Who governs the globe?* (Vol. 114). Cambridge University Press.

Beard, Alex. (2018). China's children are its secret weapon in the global AI arms race. *Wired UK*. https://www.wired.co.uk/article/china-artificial-intelligence-education-superpower.

Brown, Chad P. (2020a, December). How the United States marched the semiconductor industry into its trade war with China. *East Asian Economic Review 24* (4), 349–388.

Brown, Chad P. (2020b, September 28). How Trump's export curbs on semiconductors and equipment hurt the US technology sector. *The Peterson Institute for International Economics* (blog). https://www.piie.com/blogs/trade-and-investment-policy-watch/how-trumps-export-curbs-semiconductors-and-equipment-hurt-us.

Bughin, Jacques, & Lund, Susan. (2017, January 9). The ascendancy of international data flows." *VoxEU.Org* (blog). https://voxeu.org/article/ascendancy-international-data-flows.

Buzan, Barry, & Lawson, George. (2013). The global transformation: The nineteenth century and the making of modern international relations. *International Studies Quarterly 57* (3), 620–634. https://doi.org/10.1111/isqu.12011.

Barnes, Julian, & Chin, Josh. (2018, March 2). The new arms race in AI. *Wall Street Journal* (Life section). https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261.

Chong, Zoey. (2018, January 19). Google, China's biggest tech firm sign patent-sharing deal. *Cnet* (blog). https://www.cnet.com/news/google-tencent-china-patent-deal/.

Ding, Jeffrey. (2018a, March). Deciphering China's AI dream. Technical Report. Future of Humanity Institute.

Ding, Jeffrey. (2018b, July 16). ChinAI newsletter #19: Is the wild east of big data coming to an end?" *ChinAI Newsletter* (blog).

Ding, Jeffrey. (2019, June 7). China's current capabilities, policies, and industrial ecosystem in AI." U.S.–China Economic and Security Review Commission. https://www.uscc.gov/sites/default/files/June%207%20Hearing_Panel%201_Jeffrey%20Ding_China%27s%20Current%20Capabilities%2C%20Policies%2C%20and%20Industrial%20Ecosystem%20in%20AI.pdf.

Ding, Jeffrey, & Dafoe, Allan. (2021). The logic of strategic assets: From oil to AI. *Security Studies 30* (2), 182–212. https://doi.org/10.1080/09636412.2021.1915583.

Ding, Jeffrey, Triolo, Paul, & Sacks, Samm. (2018). Chinese interests take a big seat at the AI governance table. *New America 20*.

Dino, Gino. (2018, April 27). Google, Facebook, Amazon: Neural machine translation just had its busiest month ever. *Slator* (blog). https://slator.com/google-facebook-amazon-neural-machine-translation-just-had-its-busiest-month-ever/.

Edgerton, David. (2011). *The shock of the old: Technology in global history since 1900*. Profile Books Ltd.

Evangelista, Matthew. (1999). *Unarmed forces: The transnational movement to end the cold war*. Cornell University Press. https://www.jstor.org/stable/10.7591/j.ctv3s8qvs.

Farrell, Henry, & Newman, Abraham L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security 44* (1), 42–79. https://doi.org/10.1162/isec_a_00351.

Feigenbaum, Evan. (2003). *China's techno-warriors: National security and strategic competition from the nuclear to the information age*. Stanford University Press.

Fischer, Sophie-Charlotte, Leung, Jade, Anderljung, Markus, O'Keefe, Cullen, Torges, Stefan, Khan, Saif M., Garfinkel, Ben, & Dafoe, Allan. (2021). AI policy levers: A review of the US government's tools to shape AI research, development, and deployment. *FHI Technical Report*.

Green, Michael J. (1995). *Arming Japan: Defense production, alliance politics, and the postwar search for autonomy*. Columbia University Press.

Hassan, Hany, Aue, Anthony, Chen, Chang, Chowdhary, Vishal, Clark, Jonathan, Federmann, Christian, Huang, Xuedong, Junczys-Dowmunt, Marcin, Lewis, William, & Li, Mu. (2018). Achieving human parity on automatic Chinese to English news translation. *ArXiv Preprint ArXiv:1803.05567*.

Hersey, Frank. (2018, May 16). Lenovo founder in public backlash for "unpatriotic 5G standards vote". *Technode* (blog). https://technode.com/2018/05/16/lenovo-huawei-5g/.

Huang, Gregory T. (2004, June 1). The world's hottest computer lab. *MIT Technology Review* (blog).

Imbrie, Andrew, Kania, Elsa B., & Laskai, Lorand. (2020). The question of comparative advantage in artificial intelligence. Center for Security and Emerging Technology.

Joseph, Regina. (2019). Is China's AI future the snake in the wine? Or will our future be FAANGed? In Nicholas D. Wright (Ed.), *Artificial intelligence, China, Russia, and the global order* (pp. 199–204). Air University Press.

Keller, W. W., & Samuels, R. J. (2003). Innovation and the Asian economies. In William W. Keller & Richard J. Samuels (Eds.), *Crisis and Innovation in Asian Technology* (pp. 1–22). Cambridge University Press.

Kennedy, Andrew B. (2013). China's search for renewable energy: Pragmatic techno-nationalism. *Asian Survey 53* (5), 909–930. https://doi.org/10.1525/as.2013.53.5.909.

Kennedy, Scott, Suttmeier, Richard P., & Su, Jun. (2008). Standards, stakeholders, and innovation: China's evolving role in the global knowledge economy. NBR Special Report.

Kerr, Sari Pekkala, Kerr, William, Özden, Çağlar, & Parsons, Christopher. (2016). Global talent flows. *Journal of Economic Perspectives 30* (4), 83–106. https://doi.org/10.1257/jep.30.4.83.

Koshiba, Hitoshi. (2016). Research trends of AI based on international/national conferences proceedings. NISTEP Research Material. National Institute of Science and Technology Policy.

Lee, Alexa. (2021, January 4). Personal data, global effects: China's draft privacy law in the international context. *New America* (blog). http://newamerica.org/cybersecurity-initiative/digichina/blog/personal-data-global-effects-chinas-draft-privacy-law-in-the-international-context/.

Lee, Kai-Fu. (2018). *AI superpowers: China, Silicon Valley, and the new world order* (1st ed.). Houghton Mifflin Harcourt Company.

Lewis-Kraus, Gideon. (2016, December 14). The great A.I. awakening." *The New York Times*. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

Lucas, Louise, & Waters, Richard. (2018, May 1). China and US compete to dominate big data. *Financial Times*. https://www.ft.com/content/e33a6994-447e-11e8-93cf-67ac3a6482fd.

Mackereth, Kerry. (2019, July 19). A new AI lexicon: AI nationalism. *AI Now Institute* (blog).

MacroPolo. (2019). ChinAI: The talent. *Paulson Institute* (blog).

Malkin, Anton. (2020). The made in China challenge to US structural power: Industrial policy, intellectual property and multinational corporations. *Review of International Political Economy*, 1–33. https://doi.org/10.1080/09692290.2020.1824930.

Manulak, Michael W., & Snidal, Duncan. (2020). The supply of informal international governance: Hierarchy plus networks in global governance. *Transformation of Global Governance: Divergence or Convergence*.

Manyika, James, Lund, Susan, Bughin, Jacques, Woetzel, Jonathan, Stamenov, Kalin, & Dhingra, Dhruv. (2016, March). Digital globalization: The new era of global flows. McKinsey Global Institute.

McKinnon, John D., & Leary, Alex. (2021, June 9). Trump's TikTok, WeChat actions targeting China revoked by Biden. *Wall Street Journal*. https://www.wsj.com/articles/biden-revokes-trump-actions-targeting-tiktok-wechat-11623247225.

Metz, Cade. (2019, February 11). A.I. shows promise assisting physicians. *The New York Times*. https://www.nytimes.com/2019/02/11/health/artificial-intelligence-medical-diagnosis.html.

Murphy, Craig. (1994). *International organization and industrial change: Global governance since 1850*. Oxford University Press.

National Science Board. (2018). Science and engineering indicators 2018.

National Security Commission on Artificial Intelligence. (2021). Final report. https://www.nscai.gov/2021-final-report/.

OECD. (2017). OECD science, technology and industry scoreboard. https://www.oecd-ilibrary.org/science-and-technology/oecd-science-technology-and-industry-scoreboard-2017_9789264268821-en;jsessionid=tCTjCRrztne_C7g0saHoVjSB.ip-10-240-5-183.

Office of the U.S. Trade Representative. (2017). 2017 national trade estimate report on foreign trade barriers.

Office of the U.S. Trade Representative. (2018). 2018 national trade estimate report on foreign trade barriers.

Ostry, Sylvia, & Nelson, Richard R. (1995). *Techno-nationalism and techno-globalism: Conflict and cooperation*. Brookings Institution Press.

Ruggie, John Gerard. (1993). *Multilateralism matters: The theory and praxis of an institutional form*. Columbia University Press.

Segal, Adam, & Kang, David. (2006, March). The siren song of techno-nationalism. *The Far Eastern Economic Review*.

Summers, Tim, & Kwok, K.C. (2018, May 30). Arguments over innovation capacity miss how much the US and China are intertwined. *Chatham House Expert Commentary* (blog). https://www.chathamhouse.org/expert/comment/arguments-over-innovation-capacity-miss-how-much-us-and-china-are-intertwined.

Thiel, Peter. (2019, August 2). Opinion: Good for Google, bad for America. *The New York Times*. https://www.nytimes.com/2019/08/01/opinion/peter-thiel-google.html.

Walsh, Kathleen. (2003). Foreign high-tech R&D in China. The Henry L. Stimson Center.

Wang, Jingjing. (2016, July 18). The Whampoa Military Academy of China's internet [中国互联网的黄埔军校]. *Renwu* (人物).

Wang, You, & Chen, Dingding. (2018). Rising Sino–U.S. competition in artificial intelligence. *China Quarterly of International Strategic Studies 4* (2), 241–258. https://doi.org/10.1142/S2377740018500148.

Xu, Xiujun. (2021, February 27). The international environment and countermeasures of network Governance during the 14th five-year plan period [十四五'时期网络安全治理的国际环境与应对策略]. *China Information Security*.

Yao, Xiangkui, & Suttmeier, Richard P. (2004). China's post-WTO technology policy: Standards, software, and the changing nature of techno-nationalism. *The National Bureau of Asian Research (NBR)* (blog). https://www.nbr.org/publication/chinas-post-wto-technology-policy-standards-software-and-the-changing-nature-of-techno-nationalism/.

Zhong, Raymond, & Mozur, Paul. (2018, March 23). For the U.S. and China, a technology cold war that's freezing over. *The New York Times*. https://www.nytimes.com/2018/03/23/technology/trump-china-tariffs-tech-cold-war.html.

Zwetsloot, Remco, Dunham, James, Arnold, Zachary, & Huang, Tina. (2019). Keeping top AI talent in the United States. Center for Security and Emerging Technology.

Zwetsloot, Remco, Toner, Helen, & Ding, Jeffrey. (2018, November 16). Beyond the AI arms race. *Foreign Affairs*. https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race.

CHAPTER 44

MAPPING STATE PARTICIPATION IN MILITARY AI GOVERNANCE DISCUSSIONS

JUSTIN KEY CANFIL AND ELSA B. KANIA

Even current advances in narrow artificial intelligence (AI) have the potential to transform the world within our lifetimes. As with all transformational technologies, however, great promise is often accompanied by peril. Nowhere are the possible risks more concerning than in the domains of military affairs and national security. Increasingly, militaries worldwide are pursuing applications of AI in defense across a range of use cases, from enabling intelligence to potential advances in autonomy in complex weapons systems. At this point, whether these impacts are more incremental or transformational remains to be seen, and often the enthusiasm of militaries overtakes the state of play of AI today, given the technical limitations and challenges that remain unresolved. Nonetheless, these developments have provoked intense concerns about the risks that may arise, particularly with advances in Lethal Autonomous Weapons Systems (LAWS). To date, mechanisms for arms control and governance have struggled to adapt to these technological circumstances, and metrics to gauge progress have remained inadequate.[1]

What is the likelihood that states can achieve consensus on the core questions at stake in these debates? This chapter seeks to bring new data and methods to bear on these questions in order to evaluate trends and the content of conversations convened on AI and autonomy in weapons systems under the auspices of the Convention on Certain Conventional Weapons (CCW), which restricts the use of weapons deemed "excessively injurious" or with "indiscriminate effects."[2]

Drawing on original data, we quantify sovereign participation in CCW AI discussions. As a hypothesis-generation exercise, we use Bayesian Additive Regression Trees (BART) to evaluate a battery of factors–military capabilities, economic interests, diplomatic alignment, and regime type–in order to identify the types of characteristics associated with the most vocal states, as measured by expression wordcount. Although our analysis is strictly exploratory and does not purport to establish a theory of causation, we find that military

manpower requirements are especially predictive of a country's eagerness to participate in CCW discussions. Surprisingly, measures of military technological sophistication, such as the size of a country's drone arsenal, are less predictive, suggesting that many types of countries have interest in actively shaping the outcome, but for reasons that are probably disparate. Moving forward, we encourage future researchers to investigate the role of material and other factors in shaping international discussions.

State of the Debates

This analysis tests initial hypotheses or conventional expectations about the likely positions of states in these dates and leverages text analysis as a means to evaluate progress between 2013 and the present. The CCW has convened a Meeting of Experts on the question of emerging technologies in the area of lethal autonomous weapons systems (LAWS).[3] In 2014, the first informal meeting on the topic occurred, followed by 2015 and 2016 sessions.[4] As of 2016, a Group of Government Experts (GGE) was established, which was officially convened for the first time in 2017. The sessions continued through 2018; 2019, when guiding principles were recommended to the CCW and later adopted; and, after postponement in 2020, again in 2021 and scheduled to continue in 2022.

From the start, the UN GGE on LAWS has also been accompanied with skepticism and controversy. When even such a fundamental consideration as the definition of "LAWS" remains contested and subject to different interpretations among states, progress may remain elusive. Moreover, the endeavor of attempting to create principles and frameworks for arms control in real-time as the underlying technologies continue to advance and evolve dynamically appears at best an incredibly challenging endeavor. At the same time, the idea of a weapons system choosing and engaging its own target raises inherently complex questions of ethics and morality. The role of the human in an age of machine warfare—whether in, on, or out of the "loop" (or "kill chain" between sensor and shooter)—could change fundamentally, which can provoke fear and aversion that has manifested in the mobilization of advocacy among civil society stakeholders who have called for a ban on "killer robots" or "slaughterbots."[5] Such transactional networks can enable advocacy to influence global politics.[6]

What are these systems, and do such capabilities exist today? Among the definitions in use is that of a weapon system that uses AI to "identify, select, and kill targets without human intervention," with the decision to take life solely the purview of algorithms.[7] Among the examples raised so far to illustrate that phenomenon was an incident when drones "hunted down" soldiers loyal to a Libyan strongman as they retreated.[8] Inherently, from an observer's perspective, the level of autonomy of any drone or robotic system is challenging to verify, as such systems, whether in military or industry, can be designed to operate with varying levels or degrees of autonomy, such as within a predetermined kill-zone or based on a mission and target already determined by a human commander. There are very limited scenarios in which having no human intervention whatsoever would be advantageous. So too, while many weapons systems could be deemed "partly" autonomous, a LAWS that is *fully* autonomous in every aspect of its operations is not known to exist.

Debates within the CCW have encompassed a constellation of conflicting interests and viewpoints. At a time when military competition is intensifying among great powers that believe AI and LAWS will be critical enablers of operational advantage in future warfare, major militaries tend and should be expected to be unwilling to accept constraints on their development and deployment of capabilities expected to be or become advantageous.[9] The United States, People's Republic of China, and Russian Federation have all prioritized innovation in emerging technologies as central features of their military strategies and agendas for defense technological development. These advances are unlikely to remain limited to great powers. To the contrary, smaller and medium militaries will have the means and capacity to acquire commercial technologies that bring capabilities that were once the purview of larger and more capable forces. In this regard, current trends reflect not only military technology but also the influence of civilian innovation.[10]

As militaries and companies worldwide have worked to develop and deploy new weapons systems that hint at the future possibilities that LAWS may bring, these trends have provoked intense concerns within the civil society stakeholders concerned about the ethics and implications for human rights. From one angle, current advances in AI and autonomy are merely the continuation of prior trends that have enabled ever greater precision in weapons systems, including a degree of automation in target recognition or autonomous capability within certain operational parameters. The potential for increased accuracy could be seen as allowing for improved compliance with principles of international humanitarian law (IHL), including the imperative of distinction between civilians and combatants and the prohibition to inflict any unnecessary suffering. From a purely legalistic and technical perspective, autonomy or the introduction of AI for target recognition may result in weapons systems that are more compliant with IHL than many existing weapons systems because of the improved accuracy.

A core focus of debate is the question of human control. Ongoing advances in autonomy raise the possibility of lessening or perhaps removing entirely human involvement in critical elements of an engagement. Opponents of continued development of LAWS, especially those organized through the "Stop Killer Robots" coalition have framed their objective as "less autonomy, more humanity," aiming to combat "digital dehumanization," linking calls to #KEEPCTRL to concerns about the misuse of AI in society and inherent shortcomings in human, contextual understanding that is associated with automated decision-making.[11] These stakeholders have participated as observers and engaged prominently in UN GGE sessions on LAWS, including organizing events or demonstrations on the sidelines (and often live-tweeting sessions).

As states and civil society have grappled with these issues, questions of language, centering upon uncertainty around definitions and terminology, remains a significant impediment to progress.[12] So too, underlying these debates are complex questions about targeting processes within military organizations, the context in which human control is exercised.[13]

Beyond the divergent positions and obstacles to progress, the UN GGE on LAWS has achieved notable progress in its initial years. Encouragingly, the states participating have reached agreement on 11 guiding principles as of 2019.[14] First and foremost, the principles reaffirmed that international humanitarian law "continues to apply fully" to all weapons systems, including the potential development and use of lethal autonomous weapons systems." Beyond the relevance of law, questions around human responsibility also represent a

focus of initial consensus, given that beyond questions of control, machines cannot be held accountable in the same manner that humans can. Within a complex weapons system, the dynamics of human-machine interaction are particularly consequential, as the statement highlights. Beyond legal and policy questions, the principles also highlight best practices from a technical perspective, such as risk assessments and mitigation measures.

Despite the obstacles, the UN GGE on LAWS continues to display disparate national positions. Several countries have come out in support of a ban outright. By 2018, 26 countries were said to support such a measure. By contrast, the U.S. government has adhered to a relatively pragmatic position that centers upon safety and discussed its adherence to ethical standards without supporting a ban or binding constraints.[15] China's position has been complex and contested. In the April 2018 session, China's delegation declared the "desire to negotiate and conclude" a new protocol for the Convention on Certain Conventional Weapons "to ban the use of fully autonomous lethal weapons systems."[16] However, the definition of "fully" autonomous that accompanied this declaration in a position paper created such a narrow understanding of the weapons systems that could be covered by such a ban so as to render the rhetorical gesture of questionable relevance. In the latest iteration of China's position paper, released in 2021, highlighted principles and called for countries to strengthen regulation and research measures to link proliferation risks, while reaffirming basic guidelines around questions of human control.[17] U.S. and Russian positions as articulated in position papers also have remained relatively consistent across this time frame.

However, as of 2021, among the most recent additions to that camp is New Zealand, long a leader in disarmament, that announced in a statement in December its support for such a position.

As of 2022, the future of these debates is uncertain. In December 2021, the UN GGE convened again with limited progress and is scheduled to continue in March 2022. Given the level of interest, these debates are expected to continue within and beyond this venue.

How much progress has the UN GGE on LAWS achieved so far? To what extent have positions among participants converged and diverged on core concerns? Which characteristics are associated with a state's position and level of engagement on the topic?

Beyond the positions of individual countries, a review of the data available can allow for a more rigorous assessment of trends to date, which can inform our understanding of possible future trajectories for efforts through this venue to promote governance of LAWS. Our approach leverages extensive collection of information from conversations on disarmament, within and beyond the auspices of the LAWS GGE, and applies techniques that include sentiment analysis. On that basis, we provide further assessment of trends to date and also evaluate the factors that predict the potential pattern of life for any given installation.

Data Collection

We set out to collect, attribute, and synthesize all public statements, supplementary documents, and other comments—what we collectively refer to as "expressions" by sovereign delegations at the CCW and related UN organs—on the topic of military AI governance. Our first stop was the UN Office of Disarmament Affairs (UNODA) website.[18] The

website catalogs disarmament-oriented conversations in two of major venues, the CCW and the First Committee. There is no main directory. Users can navigate to the subdomain "/ccw-mx-xxxx/" to find UNODA's files on CCW meetings and review conferences; "/ccw-gge-xxxx/" branches link to the various GGE meetings. First Committee records are located at "/ga-c1-xx-xxxx/.[19] In all cases, documents were uploaded as PDFs.

We first used a script in the R statistical programming language to pull the website's HTML metadata on date, meeting, the submitting country, and document title, which we saved in a CSV. We then extracted the full text from the PDF files and merged this text data with its associated metadata. In some cases, the PDF had to be converted from an image into text through an optical character recognition method (tesseract). For comparability purposes, we needed all the documents to be in the same language. The majority were in English. In cases where documents were in a language other than English (several countries tended to submit in French or Spanish, and a handful in Arabic), we relied on Google's machine translation.

UNODA's available files were then cross-checked against the website of a non-governmental organization called Reaching Critical Will (RCW). RCW describes itself as the disarmament program of the Women's International League for Peace and Freedom (WILPF).[20] The organization tracks CCW meetings and, like UNODA, catalogs documents and written statements. Neither source is thought to be exhaustive. On both websites, country names are sometimes listed without any attached file. Not every country is listed for each proceeding, however. We assume these files existed but were never uploaded.

To compensate for missing documents, we also investigated CCW floor debates, where the missing submissions were presumably discussed. The UN Division of Conference Management maintains a very helpful and well-organized website.[21] Users can filter by date, keyword, and meeting series. We investigated every meeting between 2013 and 2021 with a CCW "organization" tag. There was 28GB of relevant audio data in total. We split the recordings up based on session and speaker using website metadata. After downloading the audio files, we then used a pre-trained speech-to-text model (wave2vec2) to transcribe the recordings into text.[22]

Prior to analysis, country names were standardized, and comments by session chairs, UNODA officials, NGOs, academics, and industry experts were deleted. Countries were coded by whether they are members of the European Union, the Non-Aligned Movement, the Arab or African Groups, the Shanghai Cooperation Organization, and/or a member of the US alliance network (NATO and Pacific Allies). Countries that submitted joint statements were coded as each having made that statement. In total we identified 13,255 unique expressions. Figure 44.1 maps the cumulative number of public-facing expressions by individual CCW member states through 2021.

We also merged our AI discussion data with data from four other sources: Polity Project,[23] a GDP indicator from the World Bank,[24] Andres Gannon's[25] Distribution of Military Capabilities (rDMC)[26] dataset (which compiles annual tallies of military equipment from the IISS Military Balance),[27] and United Nations General Assembly Voting Data.[28] These provide information on a number of potential correlates, ranging from economic, diplomatic, regime type, and military security. A downside is that some of these datasets do not continue through 2021. In lieu of a better method, we used the Amelia II R package to impute missing data where necessary.[29] Our assumption is that the indicators in these four datasets, all of which had partial overlap with the dates of AI discussions, would not vary

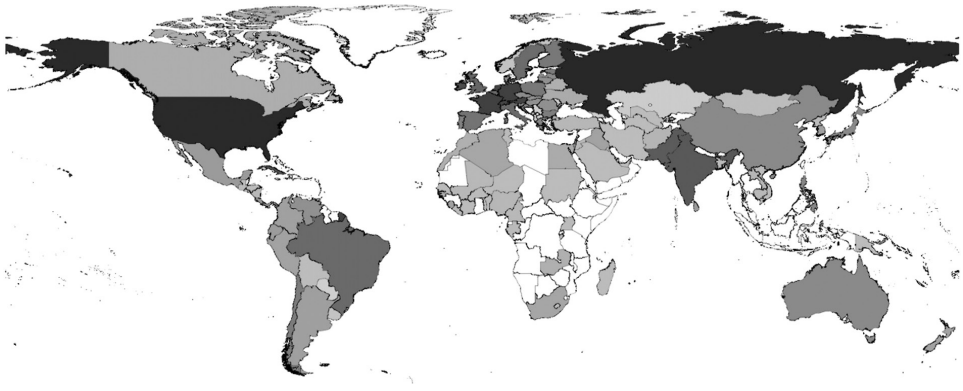

FIGURE 44.1 Number of individual government expressions by country. Darker regions indicate a greater number of expressions (from N = 1 to 549 for each country). Blank regions are non-CCW members.

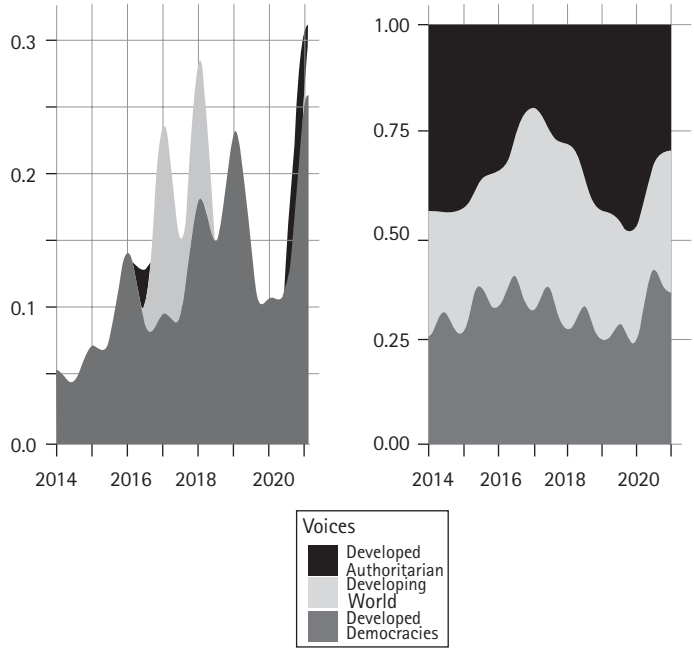

FIGURE 44.2 Density Plot (DP) showing the proportion of activity over time according to three "blocs": developed authoritarian countries like Russia and China; developed democracies like the US and UK, and the developing world. LEFT: density plots by bloc. RIGHT: relative proportion of total volume.

dramatically between their respective cutoff dates and the end date (2021). We do not impute missing text data. Finally, the data are converted to a panel format at the country-day level so each country's activity can be tracked over time. Figure 44.2 shows the proportion of commentary emanating from developing countries relative to developed democracies and developed authoritarian states.

Analysis

We wish to know what types of countries are more likely to engage in these conversations, and why. What attributes especially motivate a government to try to shape the outcome of CCW discussions on AI and LAWS?

With full text at the country-document level, our data capture both the number of times a country expressed an opinion or took a position and the total volume of what it has said. We adopt the latter as our primary dependent variable by summing up the wordcount for each expression. If certain types of countries tend to speak more often but make contributions that are less substantive, and this tendency correlates with observed or unobserved factors that explain a country's participation, looking at participation alone without weighting by expression length or complexity could very possibly bias our estimates.

What trends of national features tend to correlate with a government's verbosity on the topic of AI? Rather than advancing our own causal theory in this chapter, our objective was to weigh the various possibilities that have been advanced by other theorists. However, with more than 20 possible predictors to choose from (see Figure 44.3), we were cognizant of

	Action	wordcount	GDP	Percent_Personnel	Total_Personnel	Imports	Spending	wpm_sophistication	isr_sophistication	drone_sophistication
Action	1.00	0.42	−0.17	0.19	−0.11	0.00	−0.50	−0.43	−0.49	−0.52
wordcount	0.42	1.00	−0.03	0.07	−0.12	−0.08	−0.22	−0.21	−0.22	−0.21
GDP	−0.17	−0.03	1.00	−0.03	0.00	−0.01	0.48	0.38	0.46	0.51
Percent_Personnel	0.19	0.07	−0.13	1.00	0.16	0.08	−0.06	−0.01	−0.05	−0.05
Total_Personnel	−0.11	−0.12	0.00	0.16	1.00	0.61	0.48	0.56	0.48	0.41
Imports	0.00	−0.08	−0.01	0.08	0.61	1.00	0.24	0.24	0.21	0.19
Spending	−0.50	−0.22	0.48	−0.06	0.48	0.24	1.00	0.94	0.99	0.99
wpn_sophistication	−0.43	−0.21	0.38	−0.01	0.56	0.24	0.94	1.00	0.96	0.90
isr_sophistication	−0.49	−0.22	0.46	−0.05	0.48	0.21	0.99	0.96	1.00	0.97
drone_sophistication	−0.52	−0.21	0.51	−0.05	0.41	0.19	0.99	0.90	0.97	1.00

FIGURE 44.3 Correlation matrix of some of the possible predictors (X) of the volume of government expressions on AI governance at the CCW and related discussions (Y).

multiple comparisons problems that could threaten our ability to make reliable inferences, what statistician Andrew Gelman calls "the garden of forking paths."[30]

To minimize the risk of false positives, researchers have traditionally used methods such as the Bonferroni–Holm correction, which essentially raises the bar for detecting statistical significance. More recently, more sophisticated methods have been derived for similar purposes, such as Bayesian additive regression trees (BART).[31] The latter set of approaches provide a permutation-based alternative to parametric methods for "determining when the effect of a selected predictor is likely to be real."[32] Instead of simply down-weighting the significance of every variable in an omnibus, "kitchen-sink" regression, these approaches instead use machine learning to eliminate spurious predictors, thereby optimizing variable selection. Variables with little influence on the outcome over many iterations are discarded. While there is no substitute for theoretically grounded empirical models, techniques like BART are helpful in identifying truly significant correlations, particularly in high-dimensional situations where the number of true predictors may be small relative to the total number of possible predictors, as is the case here.

To locate an optimal subset of variables with the highest predictive value, we rely on the bartMachine package in R.[33] BART works by splitting the model into *m* trees. The terminal nodes of these trees are designed to be as homogenous as possible. Trees are compared, and those that underperform are cut down.[34] By examining which variables tend to be discarded at lower rates, otherwise known as the "inclusion proportion" (IP), the researcher can settle on an optimal set of covariates to include in the prediction model.[35] This allows us to identify variables which significantly affect the outcome over many iterations.

Again, although we do not make causal claims about the relationship between AI debate participation and its correlates in this chapter, nonparametric models like BART help us move beyond a simple "fishing expedition." To adjudicate between many possible theoretically grounded explanations for what motivates countries to participate, we suppose that BART is more appropriate than guessing at the correct model specification, particularly when many competing theories exist, not least because the researchers are blinded from the variable selection process. Another advantage to BART is that we can confirm the validity of our specification, trained on a random subset of the data, by testing its ability to predict out-of-sample patterns.

There is precedent for applying BART to social and political questions. Regardless of whether one's approach is parametric or nonparametric, the "ignorability" of treatment assignment—in lay terms, the absence of omitted variable bias—is a requisite assumption for causal identification.[36] For example, inferences may be problematic if some unmeasured feature, such as the size of a country's military, shapes both that country's ability to make itself heard at the CCW (our outcome of interest) and its interest in military AI (one of the possible correlates).[37] There are undoubtedly many potential unobserved confounders to consider, which prevents us from making any causal claims. However, we can at least make broad descriptive claims thanks to the completeness of our data collection initiative.

First, the dataset offers complete coverage of public government expressions in the principal forum for multilateral AI governance. We believe the empty links on the UNODA website were missing at random, and that the audio data we gathered effectively substitutes. Second, all UN members (and indeed some non-UN members, such as Palestine) have had the opportunity to join the CCW. Parties to the CCW have the right to speak on the record—and, in the dataset, each of them does. All but seven countries in the dataset were CCW members prior to the initiation of formal military AI discussions in 2014. Of these seven—Afghanistan, Lebanon, Ivory Coast, Bahrain, Algeria, Grenada, and Iraq (Palestine

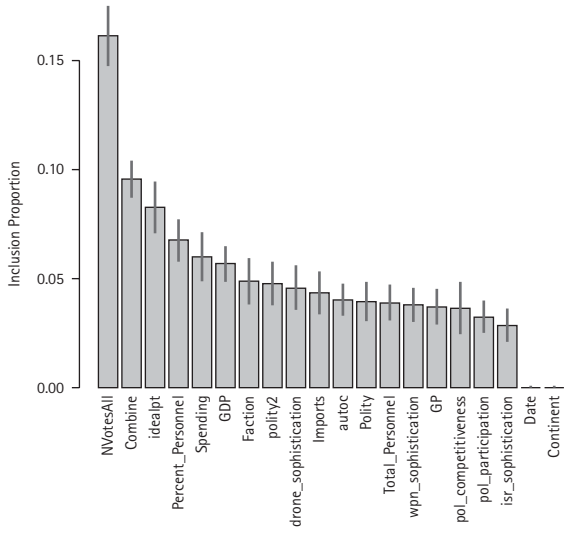

FIGURE 44.4 Inclusion Proportions (IP) for possible predictors. Higher values indicate more predictive value.

also acceded, but is excluded from the dataset)—all had signed the convention prior to 2014. Our analysis cannot definitively say whether a given characteristic X *causes* country *i* to participate in AI discussions, nor does it purport to, but the universality of coverage helps increase our confidence that observed relationships are less likely to be spurious.

Following best practices, we randomly divide the dataset into separate training and test sets. Covariate influence is explored in the training set (N = 2,545) and tested on the second subset (N = 10,710). This technique helps ensure the model is properly tuned and avoids "charges of data dredging."[38] In the training set, we first deduce the optimal number of trees by computing the lowest out-of-sample Root Mean Square Error over a five-fold cross-validation. A test for out-of-sample prediction showed that the best predictive performance is achieved at m = 50; increasing the number of trees beyond 50 does not lead to improvement. We then build a model on optimal *m* and analyze the inclusion proportions for all of our potential predictors (results are shown in Figure 44.4). Higher values indicate greater variable importance in the regression. The general takeaway from this graph—in particular its lack of a step-wise descent after overall UN voting trends (NVotesAll)—is that the ideal regressor selection portfolio is not straightforward.

Some substantive predictors do appear to stand out, however, and we can use Figure 44.2 to infer their relative weight in a country's decision to engage. This is especially the case for hard power indices. The variables with the highest inclusion proportions include how much a country spends on its military (Spending), total military size (Personnel), military size as a percent of the labor force (Percent_Personnel), and whether it belongs to an organized diplomatic bloc like the Shanghai Cooperation Organization or the African Group, denoted "Combine."

A country's overall voting activity at the UN (NVotesAll), UN alignment score (idealpt), and GDP per capita (GDP) are also predictive. Diplomatic and economic indicators may be considered proxies for hard power indices because military security both depends on and affects the economy and diplomacy. Conversely, they could also proxy for other influences, such as academic and industrial interests related to AI development—for example, whether a country is host to a Silicon Valley or a *Zhongguancun*. As a cutoff point, we select variables at the elbow in Figure 44.4 (above the 0.05 IP level) for our estimation procedure.

Next, we also explore a battery of variables related to regime type and military technological sophistication. Military AI applications are poised to make the biggest impact on command and control (C2), intelligence, surveillance, reconnaissance (ISR), and robotics.[39] We code for how many drones a country had in its military arsenal as of 2014 (drone_sophisticaton) and the number of ISR assets (isr_sophistication). We also include a total count of "exquisite" weapons such as advanced fighter aircraft, aircraft carriers, and ballistic missiles per country. Notwithstanding problems with a composite like this (e.g., how many ballistic missiles is an aircraft carrier worth? Is having one aircraft carrier better than having 20 cruisers?), we use this as a general index for how technologically sophisticated the user state is. We were surprised to find that none of these indicators, alongside the indicator for volume of weapons imports, had a high enough inclusion proportion to warrant use as a predictor in our regression models.

To capture susceptibility to civil society movements like the Campaign to Stop Killer Robots, we also included in our BART model a set of variables to measure a country's regime type (polity), level of democracy (autoc), its level of election competitiveness and voter participation rate, and whether a country is a formal US ally, US adversary, or non-aligned party (Faction). Interestingly, variables intended to capture regime type and military technological sophistication—which we might expect to have especially high predictive value—actually turned out to have surprisingly low inclusion proportions.

We do not discount the possibility that regime type or military technological sophistication may be important for the *types of proposals* different countries are likeliest to accept—for example, ranging from more prohibitive to more permissive. Because our goal is merely to measure overall engagement patterns, we do not address this question in this chapter. However, especially given the exploratory nature of this chapter, we caution against discounting these variables in future analyses.

Next, we use the six most predictive BART-selected variables (variables with an IP > 0.05) to fit three linear models with the same outcome of interest, expression wordcount.[40] Our base model is a simple linear regression (Ordinary Least Squares). We also run a second pair of linear models with one-way fixed effects (year and country, respectively). Our results are only an illustration, but since the number of words in an expression is a form of count data, future researchers might also consider Poisson or zero-inflated negative binomial models.[41] Figure 44.5 describes our results.

To preview the results, we find that two factors are most predictive of expression wordcount: the percent of a country's population serving in its armed forces (Percent_Personnel) and various diplomatic measures (UN Ideal Point, "Combine" bloc membership, and overall UN voting tendency). Interestingly, overall military spending is not predictive under any specification we explored. Likewise, GDP per capita is predictive in Model 3 (year fixed effects), but this result is not robust to alternative specifications. Models 1 and 2 also have comparatively high R^2 values, suggesting that the high-IP variables do in fact account for a substantial degree of variation in country participation, as measured in wordcount.

Interestingly, our first diplomatic index, the number of votes a country has cast at the UN only very marginally increases its verbosity in AI discussions. This is somewhat surprising. Not all 190-plus UN members are represented—a country must be a party to the CCW to participate in these talks, and only around 160 countries are—but *a priori* we might expect those that are to exhibit similar enthusiasm in both fora. UN ideal point score exhibits a slightly larger effect, indicating that the closer a country tends to vote with the US, the more it has to say about AI, on average.

	Dependent variable:		
	log(wordcount + 1)		
	Base OLS	Year FE	Country FE
	(1)	(2)	(3)
NVotesAll	0.055*** (0.001)	0.052*** (0.001)	-0.005 (0.004)
CombineEU	0.342*** (0.069)		0.598*** (0.097)
CombineLMC	-0.413*** (0.075)		-0.079 (0.121)
CombineNAM	-0.038 (0.058)		0.205*** (0.071)
CombineOther	0.093 (0.075)		0.331*** (0.094)
CombineSCO	-0.404*** (0.091)		-0.150 (0.109)
UN Ideal Point	0.074*** (0.026)	0.151*** (0.039)	0.102*** (0.038)
log(GDP + 1)	0.002 (0.006)	-2.775*** (0.344)	-0.010 (0.007)
log(Spending + 1)	0.001 (0.003)	-0.308 (0.189)	-0.005 (0.004)
Percent_Personnel	0.124*** (0.018)	0.078** (0.033)	0.062*** (0.022)
Constant	0.377*** (0.071)		
Observations	13,255	13,255	10,589
R^2	0.610	0.631	0.064
Adjusted R^2	0.610	0.627	0.063
Residual Std. Error	1.706 (df = 13244)	1.668 (df = 13123)	1.816 (df = 10571)
F Statistic	2,071.628*** (df = 10; 13244)		
Note:			*p**p***p<0.01

FIGURE 44.5 Influence of BART-selected covariates on CCW members' level of engagement in AI discussions. Regression results from four specifications.

Just as concerns have been raised that AI and robotics might affect a variety of private sector jobs, we might also expect countries whose militaries depend on a large number of personnel to take special interest in military AI governance discussions. The size of one's armed forces, or more specifically the percent of the labor force committed to military service, could positively predict participation. The reason is that, if AI can be used as a substitute for human capital, it could ease military demand on labor in such countries. Similarly, to the extent AI can make military operations more efficient, the same should be true for countries that spend more on their military, overall. Finally, countries with higher GDPs are on average more capable of building, acquiring, and operationalizing AI technologies. In short, we might face divergent expectations: wealthy, technologically advanced countries *and* smaller or less advanced countries that rely more on manpower might see equal interest in participating in talks.

While all three variables (GDP, military spending, and percent_personnel) could conceivably positively predict engagement volume, only manpower requirements (percent_personnel) does so to a statistically significant degree. We stop short of articulating a causal explanation, but this is an interesting finding that cautions nuance. Countries whose status quo military composition could undergo the most radical changes as a result of AI adoption are apparently more willing to stake positions, whereas countries that spend the *most* and *least* on their militaries are apparently equally reticent. Without a theory, we can only speculate. This could be because the latter set pools economically advanced countries who have fewer security threats, and therefore less need for military AI, with less developed countries who have *greater* security threats, but cannot as easily acquire it. In contrast, the former group might consist of countries who already spend more on their militaries *because* of a security threat, and would like to spend *less*.

While diplomatic and military manpower might be especially predictive in Figure 44.4, we caution that a country's decision about whether to engage at length could be driven not by a single, overriding condition, but rather a mix of certain military, diplomatic, and other circumstances that demand further exploration. Certainly, all CCW members see an interest in shaping the debate. Of course, a theory about why certain countries engage would require a more careful examination into whether countries engage in *prohibitive* or *permissive* ways. Absent a theory or causal identification strategy of our own, we leave a more refined investigation into subgroup effects to future researchers. We therefore encourage other researchers to use our results as a signpost for future research into the relative importance of military, diplomatic, and economic factors in engagement patterns.

Content of Discussions

What countries say is just, as well as the attitudes attached to those statements, as relevant as how often they say it. To this end, we investigate whether militarily powerful states have tended to use language that was more or less complex than the average participant. One common presumption is that more militarily advanced actors will seek to impede disarmament agenda progress by participating in a vague and noncommittal way.

To check, we ran two linear models, using the Flesch-Kincaid "readability" score of each expression as a proxy for its technical detail and specificity. Flesch-Kincaid scores are a traditional measure of how difficult a given passage is to for English readers to understand.[42] One model was a simple bivariate regression of military spending on Flesch-Kincaid score. The other included indicators for number of exquisite weapons platforms possessed as of 2014, number of drones possessed as of 2014, number of ISR assets possessed as of 2014, and

controls for UN ideal point and year. In both models we observed that militarily advanced actors actually employ language that is *more* complex.[43] Perhaps this is because the leading developers of a technology are best-suited to talk about it, or perhaps because the other side need not be specific in making broad calls for prohibitions.[44]

As a second exercise, we also conducted a preliminary sentiment analysis. Because the most obvious disjuncture in participation, although small, seems to be between militarily powerful states and everyone else, we compare expression content on this dimension. For simplicity, we leverage the Lexicoder Sentiment Dictionary (LSD) to code sentiment for each word in each expression.[45] Originally trained on newspaper content, LSD has since been used by social scientists to study sentiment in international organizations in previous literature. While we would prefer a more sophisticated method, coarse approaches such as the dictionary-based method we employ still serves our purposes of being illustrative, if not certainly robust.[46]

The text is preprocessed using the quanteda[47] and preText[48] packages in R. The preText package makes preprocessing recommendations based on how sensitive the documents are to combinatory hyperparameter changes. Numbers, URLs, symbols, and punctuation are removed. Common English stopwords are removed and the rest are converted to lowercase. Because some countries use different conventions, we also convert British English spellings to American English spellings to avoid missing duplicate words. Finally, we include 2-grams (commonly observed phrases up to two words in length). Using quanteda, we tokenize the corpus and apply a dictionary score to each token. We then convert the token list into a data feature matrix with rows sorted by date. Sentiment scores for each token are summed by document group, giving each document an aggregate score.

Figure 44.6 plots log sentiment scores for each document by time period and country type. The left plot shows sentiment among the top military spenders, and the right plot depicts sentiment for all other CCW members. Points above the horizontal dotted axis

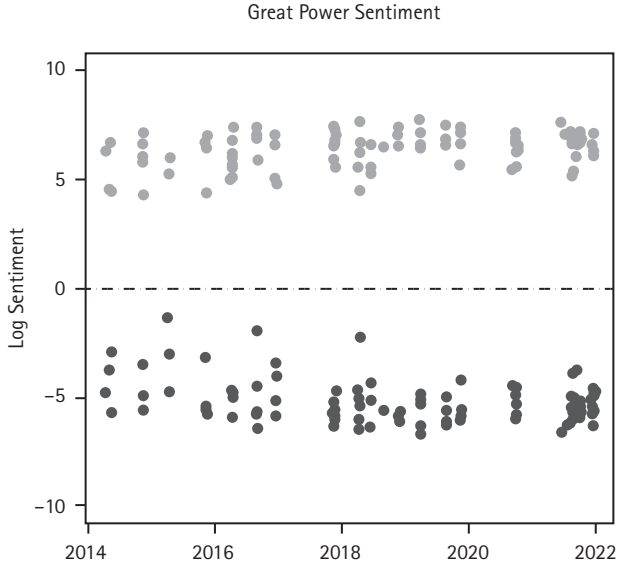

FIGURE 44.6 Sentiment scores at the document-level (logged). LEFT: "Great powers" (top-10 military spenders). RIGHT: all other CCW members. Points above the horizontal dotted axis are positive; points below it are negative. Difference is average document score.

are positive; points below it are negative. The average document score is the difference between each pair of points at the same index on the x-axis. Two findings are immediately obvious. First is that, especially for the great powers, the language used in these expressions is (according to LSD's metrics) on net more positive than negative. This could imply that such countries tend to highlight the potential benefits of AI.

The second finding is that the variance is lower for militarily powerful states than among their counterparts. While we take caution in interpreting this, one speculation is that militarily powerful states are on message, while others are of more varying opinion. In other words, AI leaders may have more in common with each other on this issue than they do with anyone else. The rest of the world, meanwhile, is not a homogenous bloc.

Using a simple word cloud, we can compare the linguistic tendencies of each group. Figure 44.7 presents such a comparison cloud. The top half (bottom half) contains words that are used by the most militarily powerful states (other states) with the total highest frequency, respectively. When CCW members address the floor, they introduce their delegation by referring to their country name. Where country names appear in this cloud, it is likely an indication that these countries tend to speak often and thus make a disproportionate contribution to the other high-frequency words in the plot. We can see that the non-aligned movement, African Group, some European states, and some Latin American states like Venezuela are represented in the bottom half, with Japan, France, the US, China, Brazil, and Russia in the top.

FIGURE 44.7 Wordcloud comparing high-frequency language of the "Great Powers" (top-10 military spenders, top) versus everyone else (bottom).

States in the bottom quadrant tend to use stronger disarmament-related language like "binding" (as in a binding treaty), "prohibitions," and "concrete... measures." They tend to prioritize "human control" (as in meaningful human control).[49] Conversely, more militarily advanced states appear to place more emphasis on "principles" (as in nonbinding principles) and "existing" (as in existing law). Hedging words like "discussion" and "think" also tend to appear more often, implying less urgency. Where states in the bottom half use "contracting," states in the top half call for "consensus."

Quanteda has a "keywords in context" function that allows us to glimpse phrasal information. We use it to observe keywords by what other types of words they are proximal to in a document. Unsurprisingly, discussion of meaningful human control is highly represented in the data. We distinguish between developed authoritarian countries like China and Russia, developed democracies like the US and the UK, and the developing world, and conduct a search of the meaningful control keyword.

As Table 44.1 shows, differences are a matter of definitional degree. Members from the developing world appear to urge immediate disarmament action. Meanwhile, developed democracies maintain that existing law is sufficient. They are careful to make distinction between human-augmented military AI and AI that is "entirely outside" human control, and stress that how military AI is used makes a difference ("predictability" in "human judgment" and "human machine interaction"). Wealthy authoritarian countries nominally agree that human control is important. Like the developing world, and unlike wealthy

Table 44.1 Developing-World, Developed-Democracies, Developed-Authoritarian

Developing-World		
force should be subject to	human control	the use of such new
law binding on all states	human control	should be defined and understood
interaction and the notion of	human control	there is general consensus among
have been used to describe	human control	such as sufficient human control or
Developed-Democracies		
suggestion that autonomous weaponry lacks	human control	and that humans, and
kill human beings entirely outside	human control	is exercised in modern militaries
the military context including how	human control	predictability human judgment and critical
understand how human machine interaction	human control	that is required to ensure
Developed-Authoritarian		
munitions must remain under the	human control	and supervision of human at
group notes that the term	human control	has not been defined and
describe human control such as sufficient	human control	or appropriate levels of human
to determine the level of	human control	that is acceptable at different

democracies, these latter countries are ostensibly in favor of a binding treaty. Yet they point out the lack of coherent definitions, arguing that further study is needed.

Discussion

This chapter tracked more than 13,000 sovereign expressions to describe participation in military AI discussions since 2014. The findings confirm many of our priors about the landscape of this debate—most notably the comparative advantages AI is expected to convey for personnel-intensive military organizations, especially among countries facing recruitment shortfalls. Yet they also caution nuance: several factors thought to be important, such as wealth (measured by GDP per capita) and military technological sophistication, exhibit less predictive power. Our analysis is admittedly coarse. All states, regardless of military or economic status, undoubtedly maintain an interest in how international AI discussions unfold. Interestingly, the disarmament agenda propagated by civil society groups like the Campaign to Stop Killer Robots seems to resonate most not with democracies per se, as existing scholarship might predict, but with the developing world.[50] However, there is also considerable heterogeneity in the opinions of lesser military powers, highlighting their relevance in these debates.

The limitations of our analysis are underscored. As a starting point, we adjudicate between different theoretical expectations for why countries might take an active interest in the military AI conversation—military, economic, diplomatic, or susceptibility to civil society pressure owing to regime type—but our analysis is not premised on a theory that we ourselves have advanced. The application of BART helps ensure that our findings are reliable by guarding against multiple comparisons, but it is by no means a substitute for theoretically motivated (and causally identified) empirical investigation. Its advantages here are best understood as an exercise in hypothesis generation, as opposed to rigorous hypothesis testing.

We also say little about the content of these discussions other than what has already been reported elsewhere, except by offering some basic statistical comparisons. We do not include any sort of topic modeling exercise, although because we collected the full text of these discussions, a future study could. More work should also be done to examine how the attitudes of different countries have changed over time. While illustrative, the sentiment analysis presented here was cursory and relied on a simple dictionary-based method. Dictionaries coded on unrelated newspaper reports in previous years may not be appropriate for highly technical AI-oriented, diplomatic discussions, so much could be missed.

While our results may not be completely surprising, we offer what we believe may be the first scholarly quantification of diplomatic discussions on military AI. There are noteworthy implications for the future of governance of LAWS. Understanding national positions and current coalitions can inform assessments of the relevance of this debate and the potential productivity of this venue for future policy outcomes. Unsurprisingly, the most advanced military powers remain the most vocal parties, lending support to the notion that the strongest parties dominate rulemaking processes for new technologies.[51] But it is precisely because the top military powers are robustly engaged in these conversations that gives lesser powers a

window of influence.[52] Potentially, the rest of the world—greater in number than core AI leaders in these fora—may yet influence the outcome if they can successfully form a coalition.

The challenges of reaching a robust consensus are no doubt considerable, yet governance advocates are not without certain windows of opportunity.[53] That said, the level of engagement in debates by a range of states and stakeholders is encouraging: sustained conversation is a necessary condition for progress. In the near term, great powers appear unlikely to accept constraints on their development of capabilities that are believed to be potentially advantageous. Nonetheless, the UN GGE can be a forum for continuing conversations in ways that can allow for learning among states and stakeholders that could create a foundation for more substantive frameworks for governance in the future.

Notes

1. Maas, Matthijs M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy 40* (3), 285–311.
2. United Nations Office for Disarmament Affairs. The Convention on Certain Conventional Weapons. United Nations Office for Disarmament Affairs, https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/.
3. https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/.
4. https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/
5. Carpenter, Charli. (2016). Rethinking the political/-science-/fiction nexus: Global policy making and the campaign to stop killer robots. *Perspectives on Politics 14* (1), 53–69; Rosert, Elvira, & Sauer, Frank. (2021). How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy 42* (1), 4–29.
6. Keck, Margaret E., & Sikkink, Kathryn. (1999). Transnational advocacy networks in international and regional politics. *International Social Science Journal 51* (159), 89–101.
7. https://futureoflife.org/lethal-autonomous-weapons-systems/.
8. https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/.
9. Ding, Jeffrey, & Dafoe, Allan. (2021). The logic of strategic assets: From oil to AI. *Security Studies 30* (2), 182–212; Horowitz, Michael C., et al. (2018). Strategic competition in an era of artificial intelligence. Center for a New American Security.
10. Verbruggen, Maaike. (2019). "The role of civilian innovation in the development of lethal autonomous weapon systems." *Global Policy 10* (3), 338–342.
11. https://www.stopkillerrobots.org/.
12. Ekelhof, Merel A. C. (2017). "Complications of a common language: Why it is so hard to talk about autonomous weapons." *Journal of Conflict and Security Law 22* (2), 311–331.
13. Ekelhof, Merel A. C. (2018). "Lifting the fog of targeting." *Naval War College Review 71* (3), 61–95.
14. "Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects," December 13, 2019.
15. Sayler, Kelley M. (2020). *Defense primer: US policy on lethal autonomous weapon systems*. Congressional Research SVC.

16. https://twitter.com/BanKillerRobots/status/984713419134853120.
17. https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202112/t20211214_10469512.html.
18. https://meetings.unoda.org/meeting/.
19. https://meetings.unoda.org/meeting/.
20. https://reachingcriticalwill.org/disarmament-fora/ccw.
21. https://conf.unog.ch/digitalrecordings/.
22. https://huggingface.co/facebook/wav2vec2-base-960h.
23. https://www.systemicpeace.org/polityproject.html.
24. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.
25. Gannon, J. Andrés. (2023). "Planes, trains, and armored mobiles: Introducing a dataset of the global distribution of military capabilities." *International Studies Quarterly 67* (4), sqad081.
26. https://www.militarycapabilities.com/.
27. https://www.iiss.org/publications/the-military-balance-plus.
28. Bailey, Michael A., Strezhnev, Anton, & Voeten, Erik. (2017). "Estimating dynamic state preferences from United Nations voting data." *Journal of Conflict Resolution 61* (2), 430–456.
29. Honaker, James, King, Gary, & Blackwell, Matthew. (2011). "Amelia II: A program for missing data." *Journal of Statistical Software 45*, 1–47.
30. Gelman, Andrew, & Loken, Eric. (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." *Department of Statistics, Columbia University* 348.
31. Chipman, Hugh A., George, Edward I., & McCulloch, Robert E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics 4* (1), 266–298.
32. Bleich, Justin, et al. (2014). Variable selection for BART: An application to gene regulation." *The Annals of Applied Statistics 8* (3), 1750–1781.
33. Kapelner, Adam, & Bleich, Justin. (2013). bartMachine: Machine learning with Bayesian additive regression trees. arXiv preprint arXiv:1312.2171.
34. For an intuitive description of how BART works, see Green, Donald P., & Kern, Holger L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly 76* (3), 491–511.
35. In the words of the bartMachine package's designers, "The inclusion proportion for any given predictor represents the proportion of times that variable is chosen as a splitting rule out of all splitting rules among the posterior draws of the sum-of-trees model. https://cran.r-project.org/web/packages/bartMachine/vignettes/bartMachine.pdf.
36. Hill, Jennifer L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20* (1), 217–240.
37. To be clear, we do include a control for the size of a country's military.
38. Green, Donald P., & Kern, Holger L. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly 76* (3), 491–511.
39. Horowitz, Michael C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies 42* (6), 764–788.
40. Results from our regressions were consistent in both the test set and the full set (N = 13,255), so we report results from the full set for increased precision. We also include a control for whether a country commented at all. Highly skewed variables (such as GDP and wordcount) are logged. As an additional step, we also tried winsorizing the data at the 95th quantile to eliminate extreme cases, although this did not have a marked effect on any of the outcomes.
41. We suspected that the data might be zero-inflated because there are many days in the dataset where no meetings were held. A dispersion test of our initial Poisson model indicated that

a Negative Binomial would be more appropriate. We also ran a zero-inflation check using the {performance} library in R, which indicated that our Negative Binomial model was underfitting zeroes, as expected. We apply a zero-inflated version of the model, using an indicator for whether a country made any comments at all as a rough proxy for the existence of a meeting. Due to space limitations, the zero-inflated model is not reported.

42. We compared with another measure, the Dale–Chall readability formula, and achieved similar results.
43. Weapons indicators were just slightly above the significance threshold (p = 0.06) but in all cases had weaker coefficients than military spending. No coefficient was especially potent, and the very low R^2 value (0.02) suggests that many other factors could also explain variation in expression complexity.
44. Canfil, Justin Key. System Shocks: Technology and Ambiguity in International Law and Arms Control. Diss. Columbia University, 2021.
45. Young, Lori, & Soroka, Stuart. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication 29* (2), 205–231.
46. Thorvaldsdottir, Svanhildur, & Patz, Ronny. (2021). Explaining sentiment shifts in UN system annual reporting: A longitudinal comparison of UNHCR, UNRWA and IOM. *International Review of Administrative Sciences 87* (4), 794–812, https://doi.org/10.1177/00208523211029804.
47. Benoit, Kenneth, et al. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software 3* (30), 774.
48. Denny, Matthew J., & Spirling, Arthur. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis 26* (2), 168–189.
49. Maas, Matthijs M. (2019). Innovation-proof global governance for military artificial intelligence? How I learned to stop worrying, and love the bot. *Journal of International Humanitarian Legal Studies 10* (1), 129–157.
50. Keck, Margaret E., & Sikkink, Kathryn. Transnational advocacy networks in international and regional politics. *International Social Science Journal 51* (159), 89–101; Price, Richard. (1998). Reversing the gun sights: Transnational civil society targets land mines. *International Organization 52* (3), 613–644; Carpenter, R. Charli. (2007). Setting the advocacy agenda: Theorizing issue emergence and nonemergence in transnational advocacy networks. *International Studies Quarterly 51* (1): 99–120; Carpenter, Charli. (2014). *"Lost" Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Cornell University Press; Young, Kevin L., & Carpenter, Charli. (2018). Does science fiction affect political fact? Yes and no: A survey experiment on "Killer Robots". *International Studies Quarterly 62* (3), 562–576; Carpenter, Charli. (2016). Rethinking the political/-science-/fiction nexus: Global policy making and the campaign to stop killer robots. *Perspectives on Politics 14* (1), 53–69.
51. Krasner, Stephen D. (1991). Global communications and national power: Life on the Pareto frontier. *World Politics 43* (3), 336–366; Lantis, Jeffrey S. (2020). *Arms and influence*. Stanford University Press.
52. See discussion of the ENMOD case: Canfil, Justin Key. (2019). Controlling tomorrow: Explaining anticipatory bans on emerging military technologies. SSRN, July 23, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3423080.
53. Horowitz, Michael C., Kahn, Lauren, & Mahoney, Casey. (2020). The future of military applications of artificial intelligence: A role for confidence-building measures? *Orbis 64* (4), 528–543.

CHAPTER 45

AI, THE INTERNATIONAL BALANCE OF POWER, AND NATIONAL SECURITY STRATEGY

MICHAEL C. HOROWITZ, SHIRA PINDYCK, AND CASEY MAHONEY

Introduction

The ability to develop and incorporate emerging technologies is an important tool of national power. States seeking to stay ahead of the curve often look for strategies using new technology to enhance their ability to assert influence on the international stage. Such strategic "ways and means" include applications in both the military sphere, where kinetic and non-kinetic technologies enhance states' coercive power, and in civilian life, where economic innovations enhance growth and productivity. Today, artificial intelligence (AI) is enabling profound advances in areas of military, health, education, manufacturing, finance, social services, and a host of other activities. Depending on how they are employed, advances in AI, as with other general-purpose technologies in history, carry the prospect of enhancing national power. AI could alter the manner in which states provide for the social welfare of their populations, signal and communicate to one another, conduct warfare, and generate wealth (Ng, 2017; Horowitz, 2018). Despite the potential for revolutionary advances in discrete applications of AI, AI may well facilitate slower and subtler, although still significant, shifts in net national capabilities (Cummings et al., 2018). Regardless of whether these changes occur incrementally or have more revolutionary impacts, they have important implications for the international balance of power.

The rapid pace of AI adoption in many spheres, however, makes the task of forecasting the effect of such applications on global power balances difficult. Variation in how national economies and militaries are using AI today exacerbates uncertainty around both the opportunities and vulnerabilities AI creates for power-seeking states (Daniels & Chang,

2021). Throughout history, the ability to mobilize and leverage national resources to exploit new technology has been a critical dimension of national power. Factors ranging from external and organizational environments, state–society relations, and ideational ethos all play a role in determining states' success in this endeavor. Risks and vulnerability specific to any technology have also affected how their adoption changes states' abilities to pursue their interests. The pervasive use of algorithms to supplement or replace roles previously occupied by humans stands to alter key features of decision-making in civilian and military life. Given these unknowns, how then—if at all—can analysts and practitioners best understand the impact of AI on the distribution of international power to arrive at optimal national security strategies?

In what follows, we show that the nascent AI-focused literature in the study of international relations (IR) alongside foundational IR concepts offer a number of ways to assess how a rapidly changing technology like AI is impacting national power. This, in turn, offers decision-makers various conceptual frameworks to use in determining their best responses in crafting national security strategy. We stop short of endorsing any particular framework or set of concepts from among those presented. Rather, our key contribution is to set a baseline for how hegemonic IR concepts and theories can be applied to the practical problems and scholarly puzzles the international proliferation of AI technologies provokes. Whether these will suffice for continued knowledge-building and problem-solving in the increasingly complex worlds of AI development and international politics is for future work to address.[1]

This chapter proceeds as follows. First, we identify the scope of our analysis and define terms. Second, we identify key pathways by which AI can both enhance and undermine national power. Then, we turn to address how these opportunities and risks are likely to change the broadest features of international politics as scholars of international security most frequently consider them: the stability of the international balance of power, the international institutional order, and the substance and compliance with norms related to the use of force. We conclude with a discussion of what current trends in state efforts related to AI suggest about the future of national security strategy and identify avenues for related research.

Scope and Definitions

Before proceeding, a few definitions and disclaimers are needed to scope the focus of our arguments. While there is no consensus regarding the meaning of an "intelligent system," AI research focuses on using machines to "supplement or complement human intelligence" in performing a range of tasks (Simon, 1995). The field of AI encompasses many areas of emphasis and methodologies, including deep learning, machine learning, robotics, natural language processing, and computer processing (Harris, 2017). AI systems can solve narrow problems, as well as adapt and learn from the environment around them to address more diverse tasks.

This delineates two broad categories of AI. Modular applications of *narrow AI* systems have narrow expertise in a particular domain and improve their performance over time

through practiced learning—not unlike a competitive swimmer perfecting her butterfly kick. Here, the machine learns, evolves, and adapts, but the hardware and software are constrained to follow a well-defined set of rules and objectives. Narrow AI applications, such as AlphaGo, are highly tailored to complete specific tasks or applications (Horowitz, 2018, p. 42; Scharre & Horowitz, 2015; Harris, 2017).[2]

General AI, on the other hand, would comprise AI systems adept at performing tasks based on parameters that have not been explicitly programmed in advance, seeking not only a range of solutions to the tasks at hand as narrow AI applications would address, but defining and solving novel tasks as well (National Security Commission on Artificial Intelligence, 2021, p. 21). Imagine, for example, an AI application that can learn to play chess, but without being told the rules of the game, and instead learns just by watching people play. As envisioned, such a "superintelligent" artificial general intelligence (AGI) system could eventually think for itself at a level that outperforms the human brain (e.g., Bostrom, 2014, p. 22).

Achieving AGI will require broad advances in hardware, software, and our understanding of cognition itself (Cummings et al., 2018, p. iv). Even something along the lines of "seed AI"—a system capable of "recursive self-improvement" through trial and error and the acquisition of new information, would require (at least initially) assistance from programmers (Bostrom, 2014, p. 29). At this point, experts disagree about how far away such advances are (Müller & Bostrom, 2014; Grace et al., 2018).[3] Recent advances in AI apply to a relatively narrow domain of problem solving (e.g., facial recognition or playing Go), require a significant amount of human planning, and are often highly brittle or vulnerable.[4] In light of the hurdles facing general AI, in this chapter, when we refer to AI, we mean narrow AI developments likely to occur in the next decades or so.

Most broadly, we conceptualize AI as a *general purpose technology* (GPT), which can be thought of as a "core invention" with the potential to significantly enhance productivity across a range of sectors, and in unpredictable ways (Cockburn et al., 2018). The electric motor, for example, brought about significant organizational and technological change in manufacturing, agriculture, residential construction, and retail sectors, and many of those changes could not have been predicted (David, 1990). As a GPT, AI is a member of a class of technologies that history has shown to hold immense potential to transform international distributions of military and economic power.

Finally, as the outcome of interest in our analysis, the concept of *power* itself merits a brief discussion. Despite the lack of scholarly consensus on its definition, power is conceptualized as comprising an actor's relational (Dahl, 1957) and latent (Bachrach & Baratz, 1962) ability to influence and control social agents in the international system. Its distribution, or "balance," depends on the preservation of the system of states and nonstate actors through the establishment of security paradigms (Claude, 1962). Power is also distributed across interlinked domains. *Military power* refers to how states use or threaten the use of organized violence to achieve their goals. Here, weapons systems serve as tools employed in the pursuit of such power. How states use such tools is often what matters in determining victory and defeat (Biddle, 2004; Horowitz, 2010). But first, states need to acquire the tools. The acquisition and maintenance of military capabilities depend, in part, on states' *economic power*. Domestic resources and political-economic control of cycles of economic innovation make possible the production of military capabilities, which in

turn reinforces economic advantage while producing a stable political order and strategic advantage (Gilpin, 1981; Kennedy, 1987).[5] Trends in scholarship that suggest the value of examining the linkages between political-economy and security affairs (e.g., Poast, 2019) motivate our analysis of the role AI could play in changing national power in each of the military and civilian political-economic sectors in the next section.

AI as a Tool of National Power

AI introduces a range of opportunities for both enhancing and exposing vulnerabilities in one's military and economic capabilities. Thus, a central question surrounding AI is how it will interact with human decision making and institutions to influence military and economic power. In the military sphere, AI can not only enhance first-order kinetic (Garfinkel & Dafoe, 2019) and non-kinetic forms of military coercion and systems of support, but also impact second-order features of militarized interactions, like crisis stability and the initiation and escalation of warfare. In the economic sphere, AI can generate wealth and improve productivity across a variety of applications—from agriculture, to healthcare, to transportation, to retail. The generation of such power relies both on novel technologies and organizations capable of integrating and using them (Horowitz, 2010; Goldman & Eliason, 2003). In other words, it is not just the technological components of AI, but also the recruitment, training, and education of those who design and utilize algorithms that matters for national power. AI is a "team sport," requiring both technical and nontechnical talent (Gehlhaus et al., 2021, p. 12). As a dual-use technology, AI also introduces the challenge of collaboration between military and civilian spheres of AI research and development. In a general sense, the successful integration of AI requires a certain degree of flexibility, communication, and openness to new approaches that is often not found in bureaucratic organizations, military or otherwise (Wilson, 1989; Halperin & Clapp, 2007).

Impacts on military power

Military power implies an enhanced ability of a state to instrumentally use violence in pursuit of its goals. The ability to credibly deter and coerce one's enemy relies on such power but is notably subjective and dynamic (Mercer, 1996). AI, like many technological advancements before it, can assist military strategists in making judgments in complex situations with imperfect knowledge (Ayoub & Payne, 2016). While narrow applications of AI will be especially relevant over the next two decades, it is worth noting that efforts to translate such applications into military power have been limited (Horowitz, 2018, p. 42; Scharre & Horowitz, 2015). If AI can increase the speed and accuracy of operations, what are the implications for the initiation and escalation of war, human control over operations, and the possibility of error and miscalculation? How does AI enhance the existing operational and tactical options at hand to achieving political aims? We will discuss some answers to these questions and many others surrounding the potential use of AI as a tool of military power.

Advantages

The use of technologies to help soldiers recognize, interpret, and even anticipate battlefield changes is arguably nothing new. In many ways, AI, like other technological advances, can simply enhance the ability of militaries to do what they already do, but faster and more efficiently. Here, AI can allow for the development of interpretive systems that learn from the environment around them and recognize complex data patterns. Deep learning algorithms, for example, observe patterns in behavior, learn, and infer. Here, combat-advising software agents could anticipate the human and natural environment and offer predictions about future attacks and other enemy actions (Jensen et al., 2020). Indeed, at the tactical level, AI could confer salient advantages when deployed effectively. In kinetic operations, automated and networked platforms can assist with the storming of enemy positions through quick coordination of maneuver and fire, as demonstrated by the prototype U.S. Marine ball drones (Ayoub & Payne, 2016; Tucker, 2015). AI software agents can also assist with information warfare campaigns, for example, by mining population sentiment through the scraping of social media data or the release of propaganda tailored to specific demographics (Jensen et al., 2020, p. 533; McGrath, 2016). Indeed, such data mining techniques can arguably enhance a state's ability to conduct political warfare.

AI could also enhance support systems, extending from intelligence and logistics to readiness and force posture. Deep learning processes allow for the recognition of unstructured information and thus the performance of a wide range of intelligence, surveillance, and reconnaissance tasks with greater speed and accuracy than human agents, offering potential informational advantages (Allen & Chan, 2017). Project Maven, for example, seeks to incorporate computer vision and AI algorithms to search through drone footage and identify hostile activity for targeting (Allen, 2017). Algorithms can also be developed for geolocating images, multilingual speech recognition and translation, and inferring a building's function from pattern-of-life analysis (CRS, 2020, p. 10). In the area of logistics and supply chain considerations, AI can be used for predictive maintenance of hardware. Real-time sensor data embedded in airplanes, for example, could be fed into a predictive algorithm to determine whether an aircraft needs to be inspected or parts repaired (CRS, 2020, p. 11). Software agents could anticipate the needs of not just individual weapons, but also the supply needs of entire joint task forces (Jensen et al., 2020, pp. 531).

Finally, AI can also be used to analyze operational tempo, weather, and the unintended effects of promotions, crew rotations, and disciplinary issues to anticipate readiness levels—the basis of which could inform changes in training prioritization, personnel rotations, and other resources (Jensen et al., 2020, pp. 531–532; Konaev et al., 2020, p. 4, 27). For example, AI could predict and notify commanders which personnel are more likely to suffer physical and psychological injuries or become too stressed to effectively perform the tasks at hand (Horowitz et al., 2018, p. 10).

AI also has the potential to power autonomous weapons systems. Such autonomy can not only help reduce the need for human supervision, but in some cases it could also reduce or even eliminate the very need for human intervention (Horowitz, 2018). Modern military applications increasingly look for forms of collaboration between humans and machines, or "combat teaming" between manned and unmanned systems (Ryan, 2017; Scharre, 2018). In this context, AI can serve as a force-multiplier that enhances capabilities at a lower human

and economic cost: a swarm of drones, for example, could potentially overwhelm a high-tech defense system (Ryan, 2017; CRS, 2020, p. 33). By enabling tactics like swarming (which requires minimal human involvement) and taking over routine tasks, AI systems may also allow for smaller and poorer countries to have larger impacts on the battlefield (Allen & Chan, 2017; CRS, 2020). The reduction of risk to personnel and the increase in tasks a given platform can perform enable a range of operational approaches to core missions. We saw this with China, in the form of a UAV "swarm assault" on an aircraft carrier (Kania, 2017).

At strategic levels, data analysis algorithms could serve as decision aids for defense leaders making choices about the allocation of resources. AI can assist command and control with managing the complexity of the operational environment, enhancing situational awareness and, thus, reducing vulnerability to fratricide and an adversarial attack (Lewis, 2018, p. 7; Konaev et al., 2020, p. 24). Such improvements to the quality of information available to decision makers and their strategic readiness could offer a significant advantage on the battlefield. AI can allow militaries to perform activities that usually require the efforts of many people with just a few people, or even no people at all. In that way AI can reduce the costs of operations, while at the same time provide the benefits of precision and predictive analytics (Allen & Chan, 2017).

Challenges

As one might expect, however, such capabilities introduce some challenges. For example, some new situations might not match those previously observed by an AI system, exacerbating the tension and uncertainty that undergirds elite decision making when making key security decisions (Allison, 1969). AI systems are inherently capable of actions that are—to varying degrees—unpredictable. Such unpredictability may result in miscalculation of an adversary's actions and inadvertent escalation (Altmann & Sauer, 2017). Indeed, it is difficult to assess an adversary's as well as credibly demonstrate one's own AI capabilities. Such capabilities are arguably less measurable than other forms of military power, making it difficult to clearly and accurately communicate the degree to which AI systems may extend one's military power: the coercive effect of AI may be limited by the ability for states to demonstrate their capabilities, inviting opportunities for miscalculation and deterrence failures (Davis, 2019, p. 15; Wilner & Babb, 2020).

AI also provides systems with the ability to react quickly. Such acceleration, however, may increase the pace of combat in a destabilizing and potentially destructive manner, especially in the event of a loss of system control (Scharre, 2016, p. 35). The increased operational speed may create an incentive to strike first, putting the defender at a disadvantage and needlessly accelerating and thereby intensifying a conflict (CRS, 2020, p. 38; Scharre, 2018). AI-driven ISR, for example, could reduce the time available for the management and avoidance of crises, resulting in rapid decision making and the execution of poorly planned operations (Davis, 2019, p. 10). Moreover, fast-paced operations will require coordination and communication across varying chains of command, organizational cultures, and levels of clearance (Davis, 2019, p. 12). The prospect of losing human control introduces indiscriminate and unpredictable effects, conflicting with principles of distinction, proportionality, and precaution inherent in humanitarian law (Gill, 2019, p. 117).

Estimates of strategic intention and power often rely on assumptions that simplify complexity and reflect social and organizational biases. For example, military analyses often consider tank and bomber counts rather than geographic limitations and logistical footprints (Marshall, 1966). Such assumptions could arguably still drive AI algorithms and thus result in a reduction of errors: "such as counting tanks as opposed to counting tanks recently repaired, operated by a trained crew, and optimal for a contingency in the open desert" (Jensen et al., 2020, p. 540).

Biases concerning race, religion, and gender can also be inserted into AI systems. Big data and algorithms can reinforce and amplify existing inequalities (O'Neil, 2017). For example, consider the use of algorithms for hiring that rely on information concerning the relationship between employee attributes and job performance (Tambe et al., 2019). In 2018, Amazon discovered that its hiring algorithm was giving higher scores to male applicants because it was built on historical job performance data when most of the employees were white men. As a result, the scores for female job candidates were discounted (Meyer, 2018). In the military sphere, assumptions surrounding the age and sex of civilians may inform not only personnel decisions, but also calculations for targeting: the civilian casualties following drone strikes are often the result of miscalculations coming from the assumption that all military-age males in the blast-area are combatants. By focusing on women and children as civilian victims of war, civilian men are often victims of direct violence (Carpenter, 2005). These concerns are timely: given that AI systems are at early stages of development, the industry responsible for its creation risks interesting prejudices that may influence decisions for many years (Garcia, 2017).

Impacts on economic power

AI could also impact the balance of power indirectly through shifts in economic power that influence overall national power, and in that way have longer-term consequences for military power. Robotics, sensors, and their connection through digitization could shape the kinds of activities an economy pursues.[6] Applications ranging from the AI system DeepStack beating humans at the Texas Hold 'Em poker game (Caruso, 2017), to Amazon's cashier-less grocery stores (Wingfield, 2018), to the creation of virtual assistants such as Alexa and Siri (Madrigal, 2011) have highlighted both the capability of technology to boost economic growth, as well as to perform many of the functions that previously required a human worker (Frey & Osborne, 2017; Executive Office of the President, 2016; Furman & Seamans, 2019). Increased investments and advances in AI have compelled economists to the following questions about AI: (1) Does AI increase productivity by reducing the amount of labor needed to produce goods; (2) Does AI take away jobs or lead to more jobs; and (3) Does AI lead to increases in inequality? Recent research suggests that AI has the potential to increase productivity growth but may have mixed effects on labor demand, particularly in the short run. In particular, some occupations and industries may do well while others experience labor market upheaval (Furman & Seamans, 2019). These findings have important implications when considering whether or not AI can enhance a state's economic power.

Productivity

When considering whether AI will foster productivity growth, scholars often draw comparisons to advances in automation throughout history. Examples ranging from the

steam engine (Crafts, 2004) to the electrification of manufacturing (Rosenberg, 1983, 1998; Schurr et al., 1990), to information technology (Oliner et al., 2007; Jorgenson et al., 2008; Bloom at al., 2012) suggest that the ability to perform tasks more quickly and efficiently and with less labor will contribute to economic growth. As a GPT, AI enables a range of follow-on innovation (Cockburn et al., 2018). However, the commercialization of innovative ideas that build off of the technological progress of AI relies on complementary investments, which often take time (Brynjolfsson et al., 2019).[7] For example, the productivity impact of electric motors was largely driven by increased flexibility in the location and design of factories (David, 1990).

Nonetheless, research on robotics indicates that such advancements have generally had a positive effect on productivity: for example, Graetz and Michaels (2018) find that robotics added 0.36 percentage point to annual labor growth between 1993 and 2007 for the 17 countries in their sample—a similar magnitude to the effect of steam engine in the United Kingdom.

Jobs

Throughout history, automation has been a substitute for human labor. This has resulted in a great deal of concern that technological advances, such as improvements in the processing power of computer chips, would cause irreversible damage to employment opportunities and wage rates. However, while this has probably been true in the short term, it has often resulted in the creation of complementary jobs in the long term, with varying effects across occupations (Autor, 2015). For example, cab-horsemen became taxi drivers who, in turn, became "partner-drivers" for ride-sharing services such as Uber and Lyft (Cummings et al., 2018, p. 36). Such effects suggest that workers will need to learn new skills to remain employed at their current job or transition to a new area of employment. Acemoğlu and Restrepo (2018) argue that while AI will displace roles previously occupied by humans, such displacement will also result in growth by creating new types of jobs that had never existed. Thus, in turn, AI will have a "expansion effect" as new tasks are created in which labor has a comparative advantage over machines.

However, the use of AI and the treatment of workers are contingent on management practices and techniques (Brynjolfsson & Hitt, 2000). In other words, whether or not rapid shifts in AI will result in unemployment across large segments of the population may depend on whether or not companies seeking to integrate AI across their organization invest in the retraining and compensation of workers for new roles requiring a different set of skills (Goolsbee, 2019; Furman & Seamans, 2019, p. 163).

Inequality

However, other economists note that while AI may produce positive long run prospects for jobs and overall growth, it may increase inequality. Where AI replaces unskilled workers and complements skilled workers, the wages for unskilled workers will likely decline (Furman & Seamans 2019, pp. 171–172). While highly skilled workers are often able to learn new skills when technology automates aspects of their jobs, this is more challenging for less skilled workers (Agrawal et al., 2018a, p. 19). For example, when computers reduced the time accountants needed to spend adding columns of numbers they learned a new set of skills, leveraging the efficiency of computers (Agrawal et al., 2018b). On the other hand,

manufacturing workers have struggled to adjust to the automation of factories (Agrawal et al., 2018a, p. 19). By increasing inequality across worker groups, AI introduces a distributional cost and may be excessive from a welfare point of view (Acemoğlu, 2024).

Patterns of diffusion and dual-use applications

Given that AI can be used for both civilian and military purposes, the kinds of tools available to a state interested in asserting influence in the international system will depend, in part, on whether the systems come from military or non-military institutions (Horowitz, 2018, p. 39). Today, AI innovation is largely driven by the private sector. Moreover, commercial interest in AI has created more promising career opportunities for talented engineers than that of military organizations, including higher salaries and benefits (Horowitz, 2018, p. 50). Some scholars argue that if AI remains a primarily commercially-driven enterprise, AI capabilities will diffuse more rapidly around the world, diminishing "first mover advantages" and potentially altering the balance of power: the period in which AI innovators would enjoy a market advantage "shrinks when countries or companies can acquire or copy others' advances relatively easily" (Horowitz, 2018, p. 51). Indeed, the commercialization of AI suggests that traditionally weak states and non-state actors may have the opportunity to purchase and use the technology, leveling the playfield with more powerful actors and potentially opening up new avenues for coercion of much stronger adversaries (Wilner & Babb, 2020, p. 407). In this scenario, militaries are likely to focus on amassing a large number of narrow AI systems (as opposed to fewer, more advanced systems).

On the other hand, military applications of AI which are not dual-use and based primarily in military research are likely to be harder for other countries to mimic. Military applications of narrow AI, such as battle management algorithms to aid with the quick coordination of operations, do not always have straightforward civilian counterparts. The adoption of such platforms could generate significant first mover advantages for wealthy companies and countries who are able to adopt higher-end AI capabilities earlier on (Horowitz, 2018, p. 39, 52). However, adoption will also generate significant organizational challenges: great powers who are able to exploit costly advances in AI are also the most likely to struggle to use organizationally disruptive technologies (Horowitz, 2018, p. 53; Gilpin, 1981). AI systems will inevitably vie with the entrenched preferences of defense establishments: the bureaucratic structures and parochial interests of those within militaries will influence how AI is adopted and used to produce military power (Snyder, 1984; Wilson, 1989; Halperin & Clapp, 2007). Here, military competition would be largely concentrated among existing great powers and the battlefield advantages allocated to the militaries more open to replacing existing platforms and approaches with novel yet disruptive algorithms (Horowitz, 2018, pp. 53–54).

Data access, ownership, and intellectual property protections are also a concern when considering how AI technology may diffuse across the international system. Machine learning relies on data to make predictions. The availability of massive datasets (and the powerful computing hardware needed to process them) has played an important role in current advances in AI (Horowitz et al., 2018). Thus, larger companies and countries with more data may get the most benefit from AI. However, the collection and use of such data also introduces a range of privacy concerns: data (1) is relatively cheap to store and thus

may last longer than intended or desired; (2) can be repurposed for different uses; and (3) may contain information about many individuals, regardless of who creates it (Agrawal et al., 2018a; Tucker, 2019). Thus, AI datasets have the potential to harm individuals who are unaware of how their data is being collected and used. Regulations such as the 2014 "right to be forgotten" and the 2018 General Data Protection Regulation, for example, demonstrate a stronger stance on protecting the digital rights of individuals (Meltzer, 2020). As privacy regulation becomes an increasing concern to policymakers, it is worth noting that such regulation may also impact AI innovation and diffusion (Agrawal et al., 2018a; Goldfarb & Tucker, Forthcoming). Here, countries with more lax privacy policies may have a short-term advantage, creating a "race to the bottom" in which countries seek to get ahead of each other in AI by prioritizing data collection over protection: countries with stricter privacy policies may have a disadvantage when it comes to accessing the data required to build many types of AI (Agrawal et al., 2018a, p. 10).[8]

The governance rules and policies surrounding the collection and use of data also has implications for military applications of AI. Deep learning algorithms depend on the training data they are created from. As a result, the initial application area of an AI system can have a persistent innovation advantage through their control over data. Thus, the balkanization of data within a given sector could inhibit innovative productivity in the other (Cockburn et al., 2018, p. 15). Given the highly dynamic conditions of the battlefield, a lack of sufficient data could produce unreliable and vulnerable AI systems in the military sphere (Horowitz & Scharre, 2021). Here, data concerns extend beyond access to problems with traceability and the interoperability of data collected by difference systems (Tarraf et al., 2019, p. xiii). Thus, the systematic collection, curation, sharing, and strategic guardianship of data across civilian and military spheres has important implications for how AI tools are designed and implemented.

AI and the International System

In light of the multiple pathways by which AI applications are likely to affect the power of state actors in the international system, IR theory suggests an equally diverse set of frameworks in which to think about how these changes will affect features of international politics. Here, we consider how dominant concepts and theories in the IR literature suggest conceptualizing AI's effects on three features of the system: systemic polarity (i.e., the balance of power), institutionalization and international governance, and the substance of international norms.

Ongoing shifts in each of these areas complicate efforts to forecast AI's potential impacts. China's rise, the continued influence of Europe and Russia, and India's longer-term development are transforming the interstate system to a multipolar one (Mearsheimer, 2019) where neorealist IR theory expects greater instability to arise (Waltz, 1979). Symptoms of anti-globalization backlash (Walter, 2021) have served to erode confidence in patterns of increased interdependence that the international institutions of the post-World War II order helped create (Keohane, 1984). Not least, these changes call into question the normative bases on which the contemporary international order is founded (Ruggie, 1982; Lake et al., 2021). Understanding how these trends interact with the emergence of AI will require

analysts and policymakers alike to adopt clear concepts and theories as a basis for understanding the relationship between AI and international politics. As such, the following discussion provides a baseline of how current theory might illuminate this relationship as a first step toward evaluating whether new concepts and theories are needed.

System structure and the balance of power

A baseline feature of international systems, analyzed at the global or regional level, is system structure. Whether advances in AI can impact system polarity depends, most fundamentally, on the contribution AI makes to the relative military and economic power of a given system's constituent members. Both "raw" AI resources (e.g., large quantities of raw data, an AI talent pool, access to computing) as well as domestic political and institutional factors (e.g., military organizations and domestic political-economies capable of adapting AI technology to applied uses, public–private partnerships, and political will) are required for "AI great powers" to exist (Horowitz et al., 2018). The United States and China are currently leading in AI research, but other countries, including Russia, Israel, France, and the United Kingdom, among others, are also investing heavily in military and economic applications of AI.

Yet, because the identities of the states who, in the longer run, will prevail at increasing national power using AI the most remain undetermined, it is useful to ask how AI might contribute to the stability of the current balance of power in the international system or, alternatively, induce change. Whether AI reinforces today's multipolar distribution of power depends most on how its adoption, particularly in military applications, affects two factors key in IR literatures on the balance of power.

First, AI may change the speed and relevant time horizons over which power shifts among states occur. Transitions among hegemons in interstate systems are often accompanied by great-power wars (Organski, 1958, Gilpin, 1981), which can serve to maintain a status quo or change the system's polarity. Commitment problems that inhibit rising powers from credibly promising not to exploit others also arise when the technologies to which states have access generate first-strike incentives (Fearon, 1995; Powell, 2006). The multiple effects, previously discussed and in Kania and Canfil (this volume), that operational applications of AI technologies could have, primarily through speeding the pace of battle, could exacerbate first-strike incentives in a crisis involving AI tools (Altmann & Sauer, 2017; Horowitz, 2019). Uncertainty around what impacts AI will have on military and economic power also presents information problems likely to complicate state efforts to assess one's own position vis-à-vis others and react rationally. For instance, the information problems likely to plague early debuts of such technology in militarized interactions could further destabilize the crises that can accompany systemic power transitions (Altmann & Sauer, 2017; Horowitz, 2019). Beyond the crisis setting, states also will struggle to know with certainty how AI is changing the value of the assets they already possess or of the utility of the strategies in which they have invested to exploit them (Daniels & Chang, 2021). States' perceptions that novel applications of AI in military operations might give them a strategic edge could also incentivize longer-term pursuits of AI-enabled military power (Talmadge, 2019)—to their benefit or peril. Together, these factors may exacerbate the commitment problems of longer-term systemic instability more acutely, particularly in cases in which it is unclear

which of a global or even regional powers' capacity to organize the adoption of AI into military and economic power is greatest.

Second, like other innovations in technology and military doctrine, AI may change the incentives states have to adopt given security strategies. Consider how states may internally balance, by building arms (Herz, 1950; Jervis, 1978), or externally balance, through alliances, against powerful and threatening actors (Waltz, 1979; Walt, 1987). The continuities or changes in alliances that AI may prompt not only determine the composition of international coalitions that compete at the global level. Transformations in the identities of the membership of alliances (Snyder, 1997), hierarchies (Lake, 2011), and great-power condominiums (Braumoeller, 2008) that largely rest on the international system structure also hold the potential to generate new sites of political disputes over alliance aims and strategy related to AI. The features of AI technology present significant technical obstacles for states who wish to integrate AI-enhanced military technologies (Lin-Greenberg, 2020). The technical challenges of AI integration and subsequent political risk of becoming overly dependent on the data, algorithms, or hardware produced by others could incentivize some states to rely less on allies for their security by pursuing independent means of security, including by developing their own AI systems. Although some fear this could lead to arms racing dynamics (Geist, 2016), a scholarly consensus is building that "arms racing" is an inappropriate concept to describe developments in AI (Roff, 2019), or at least without greater precision (Scharre, 2021). On the other hand, solving problems related to integration challenges between allies and partners could lead to new innovations or help allies use choices in the technical design of national AI systems to signal their political intent (Imbrie et al., 2020). As such, states' choices over balancing strategies and AI adoption are interlinked.

International order and security institutions

The constitution of dominant international orders, defined as the set of institutions that govern the interactions among actors in the international system, has been closely linked through modern world history to technology development and the ability of powerful states to develop and maintain the organizational capacity to effectively use technology (Ikenberry, 2001). The predominance of mercantilist interests represented by Spain, Great Britain, and other seafaring powers from the sixteenth century related to those states' edge in naval technology. In the nineteenth century, the threat of the mass army sustained the competitive balancing act of nationalist and conservative interests that national leaders played in the Concert of Europe. And, in the Cold War, nuclear weapons enhanced the stability of the bipolar system that allowed the liberal and communist orders to develop international rules within and among their respective spheres. How will AI change international order? Many of the other contributions in this volume address the ways in which the advent of AI has prompted decisions of international governance in economic and human rights/data protection spheres. Here, we consider the pathways by which AI advances may influence the international security order through its impacts on international security institutions—from issue-focused treaty arrangements and developments in international law, to activities under the auspices of standing international organizations.

The set of international security institutions most relevant to the advances of AI are those related to arms-control treaties, codes of conduct, and verification bodies that might govern the military systems that use AI. Early work to specify the requirements of systems to verify "whether an AI exists in a system or subsystem and if so, what functions that AI could command" suggests that while feasible in principle, a plethora of concerns about spoofing and how to organize technical solutions in broader verification mechanisms have yet to be addressed (Mittelsteadt, 2021). Parallel challenges notwithstanding, resolving complex verification tasks to make multilateral arms control agreements technically possible, in the past, has taken decades of international scientific cooperation to address, as with the Comprehensive Nuclear-Test-Ban Treaty (Melamud et al., 2014). Maas (2019) explores how recent steps toward creating epistemic communities of actors committed to regulation and institutionalizing transnational norms surrounding AI suggest how an AI arms-control regime could develop along similar pathways as in other domains of international governance (Finnemore & Sikkink, 1998).

Accordingly, many actors in the international community have focused efforts in this regard on high-profile political concerns related to potential for the application of AI in lethal autonomous weapons systems (LAWS). States have put forth concerns about the regulation of LAWS for debate under the Group of Governmental Experts (GGE) on Convention on Certain Conventional Weapons, although consensus about how international legal instruments ought to regulate LAWS has eluded the GGE. The reasons for this range widely, from disagreement about whether existing laws of armed conflict suffice to ensure basic protections (Reeves et al., 2020), to a lack of clear definitions about technology that militaries have yet to deploy (Ekelhof, 2019).

The many applications of AI other military systems that are non-lethal or non-autonomous, could give states ample room to explore how AI-specific institutions like codes of conduct or confidence-building measures could reduce the potential for unanticipated consequences and accidents involving AI-enabled systems (Horowitz et al., 2020). In any institution that regulates or controls military applications of AI, states will likely want to maintain sovereign authority to use AI in critical national security applications. Nevertheless, the breadth of countries and organizations, from the United States, China, and India, to the OECD and EU, who are adopting AI ethics guidelines (EU, 2019) suggests shared interests that states may seek to institutionalize in the future. These interests, at the heart of how states may seek to shape the international security order, are sure to interact, as we consider next, with how transnational ideas about the appropriate use of military applications of AI evolve.

International security norms

Last, we ask how the development and proliferation of military applications of AI may impact the predominant norms that regulate the militarized interactions of states today. Norms, or behavioral standards ascribed to actors with certain identities (Katzenstein, 1996), function as institutions by setting social rules that, even if only implicitly, raise the prospects that actors who break them will be sanctioned by others who acknowledge them as legitimate. As such, norms shape how actors see the appropriateness of certain courses of action available to them (e.g., March & Olson 2011). Given that the advent of AI is only

beginning to open *new* possibilities for such action in the realm of military strategy and the politics of international security, how important actors in the international system will come to understand the legitimacy of the use of AI tools in military settings remains unclear. Which norms surrounding military applications of AI are likely to emerge, and how? Most broadly, international norms often emerge through a process by which the individual action of norm entrepreneurs (Finnemore & Sikkink, 1998) leads others to argue about the correctness of competing perspectives (Risse, 2000) that leads to the spread ("cascade") of a norm that states internalize and act upon.

The breadth of actors engaged in debates about the proper use of AI, in general, suggests that norm entrepreneurship is in no short supply. AI applications in high-stakes social decision-making settings, like the criminal justice system and medicine, have brought the ethics of AI to the attention of a broad audience of advocates; scholars, experts, and engineers; policymakers; and the general public (Crawford, 2021). States, whose use of AI in military settings international norms might regulate, are responding by adding their views to these debates. Beyond the industrialized states who are leading in the development of AI, states of the Global South have taken a leading role in some areas, particularly in discussions about limitations on autonomous weapon systems (Bode, 2019).

Whether regulative norms limiting which kinds of AI systems states ought or ought not to employ in any (or in certain) contexts emerge will necessarily depend on whether the arguments such entrepreneurs make succeed in persuading states with the ability to develop and use AI in such ways. Thus far, across both democracies and autocracies, those with the greatest capacity to adopt military applications of AI have not decided to join the ranks of others calling for a ban on autonomous weapon systems (Rosendorf, 2021). Whether this changes in the future is likely to depend on changes in the *context* in which entrepreneurs seek to persuade possessor states or in the *strategies* they employ to do so. For example, norms against autonomous or other classes of AI-enabled weapons might follow the footsteps of the nuclear taboo that arose as a consequence of the context of a catalyzing event of catastrophic use created in which normative arguments took on new meaning among decision-makers and the public (Tannenwald, 1999). Even without such a catastrophe, opposition to LAWS in global public opinion appears to be steadily increasing (Ipsos, 2019). Alternatively, advocates might also choose to frame their arguments for stronger regulation of military AI in ways that depend less on the potential ways in which autonomous weapons may *violate* international humanitarian law (IHL). While IHL-based arguments helped develop negative norms *prohibiting* blinding lasers and other indiscriminate weaponry, the unique characteristics of AI-based weapons may create the opportunity—or need—for advocates to rely on discursive approaches that emphasize the need for positive norms *prescribing* human control (Rosert & Sauer, 2021) or a focus on human rather than national security (Roff, 2018).

If regulative norms do gain broader legitimacy as military applications of AI proliferate, the question of how they will shape state behavior remains. Of particular salience is whether states who adopt AI technologies in lethal applications, as opposed to those who do not, will comply with broadly held conceptions of what constitutes appropriate use of such technologies. Early efforts by the EU to institutionalize normative standards for ethical AI in its General Data Protection Regulation (GDPR) appear to reflect realist predictions that states will tend to disregard the relevance of international norms when it comes to the

defense of core national security interests. Rights to explainable AI and data protection find their limits in the national security "cutouts" that Article 23 of the GDPR provides, whereby EU member states may restrict individual data-subject rights for national and public security, defense, and related purposes.

On the other hand, some AI leaders have taken steps that seem to respond to public pressures to respect, if not simply define, nascent norms that might govern military applications of AI. In September 2020, the U.S. established an AI Partnership for Defense (PfD) among "like-minded" states adopting military applications of AI to strengthen practical cooperation on data-sharing, interoperability, and other areas. Tellingly, the first meeting of the PfD focused on "ethical principles" to guide further AI-development efforts (U.S. Department of Defense, 2021). To be sure, dominant global military powers may be reluctant to comply with more restrictive international norms that non-state or low-AI-capacity states may continue to champion in the future. At face value, however, the broad policy frameworks that states with the most powerful militaries, the U.S. and China included, have established with reference to international standards and ethical concerns provide reason to believe that international norms and state behavior will continue to coevolve in dialogue with one another.

Conclusion: National Security Strategy in an AI World

AI has the potential to both enhance and undermine relative national power, introducing significant uncertainty for state actors. This chapter addresses a range of potential concerns, as well as advantages for states interested in enhancing their ability to influence others and stay ahead of the curve. Applying fundamental concepts from the study of military and economic power, patterns of international technological diffusion, and systemic IR theory, we have sought to illuminate some of the many questions policymakers and scholars will find of interest in their applied and theoretical pursuits. We conclude by taking stock of where states' efforts have been most concentrated in pursuing their national interests in a competitive and increasingly AI-laden international security environment, and by identifying avenues for future scholarly work that can inform the national security strategies states will adopt in the future.

First, most states remain in the early stages of thinking about how AI will be integrated across civilian and military spheres and how such changes can be harnessed, particularly in the long term, to enhance their ability to achieve their national objectives. To this end, states are releasing national policy plans, integrating considerations for AI into grand strategic, national security strategies, and economic planning documents. In a similar vein, states have also embarked on a range of collaborative undertakings to explore opportunities for international cooperation where shared interests exist. A range of new organizations and multilateral fora are discussing the risks of "militarizing" AI, helping to shape states' thinking about not only the security costs and benefits of doing so, but also the ways in which they may affect perceptions of the legitimacy and appropriateness of their pursuits of such technology. The ongoing evolution of discourses and decision-making processes

around AI at the state- and interstate-levels presents scholars with good opportunities to describe and conceptualize these phenomena.

Second, beyond states' efforts to forecast and strategize decades into the future, states are also making near-term investments to hedge against possible first-mover advantages that states like the United States and China could gain from AI. For some other states, who may face fewer external security threats or have less ambitious goals in international politics, the imperative to compete at the leading edge of military applications of AI is less dire. For them, the payoffs of potential "second-mover" advantages are alluring. The real power of AI for states may derive from the extent to which users can effectively pool data in shared architectures and leverage decision-making tools in national security organizations. So, for states whose militaries are likelier to be military AI "takers" (i.e., from technology exporters or powerful allies) rather than "makers," targeting investments in applications proven elsewhere could be a more efficient, less cost-intensive path to retaining power in an AI-laden security environment than the broad-based R&D efforts others are pursuing. Success in this approach is likely to require cooperation with technological leaders and careful efforts to transfer tacit knowledge and tailor organizational solutions to recipient states. Here, as investments help spur new innovation in the many sub-domains of AI across the world, scholars have ample space to develop and test theories of the factors that encourage, optimize, or otherwise affect the proliferation of such technologies and of the multiple potential effects they can have on national power, as this chapter begins to develop.

Third and last, states' efforts to predict and manage the way AI is affecting their security interests will continue to be complicated by the speed with which innovators in the private sector are speeding ahead of regulators' ability to keep up. Hyperglobalization continues to facilitate rapid technological diffusion and the development of new tools outside the strictures of state-led R&D. How states choose to organize their domestic political-economies in an AI world, as we have suggested, will also have consequences for their success or failure to maintain the power to influence the terms upon which they compete with others in security affairs. So, in many ways, states' responses to the domestic political challenges AI innovations present will play a crucial role in their position internationally. Scholars have responded in earnest to the various challenges and political upheavals associated with the processes of globalization and these processes' interactions with domestic politics; the breadth of scholarship on AI governance reflected in this volume shows that this research agenda continues. For states and the individuals they govern, participation in the global AI ecosystem offers benefits in innovation and security, but it also opens up risks of exploitation. Scholarship can help identify the conditions under which states are likely to find a balance between openness to innovation and security and, in turn, inform the decisions that will determine how AI shapes the global balance of power in the decades ahead.

Notes

1. Well-known critiques of the changing, if not declining, relevance of the state vis-à-vis non-state actors in contemporary international politics (e.g., Strange, 1995) may lead some to call into question the utility of what follows as an essentially state-centric analysis. We embrace these critiques and offer this chapter as a baseline from which to expose the specific

potential or poverty of this approach to interrogating the effects of AI in international politics.

2. AlphaGo is a game-playing algorithm created by the AI company DeepMind. In March 2016, it defeated world-champion Go player, Lee Sedol, four games to one, overturning existing wisdom on Go game play (Metz, 2016).
3. Guesses range from seven years to 70 years or never (Market for Intelligence Conference, 2017).
4. Moreover, AI developers still struggle to develop systems that can demonstrate human-like capacities for processing semantic understandings of concepts. For instance, Google's learning algorithm for identifying cats in YouTube videos was programmed with unsupervised methods but could not demonstrate understanding of what a cat was (Ayoub & Payne, 2016).
5. Notably, this understanding of power, conventionally favored in IR, is one of "power over." It is a top-down and coercive understanding of social reality (Runyan & Peterson, 2014, p. 101). While this chapter does seek to incorporate dimensions of institutionalized bias and inequality into our understanding of the role of AI, it is worth noting that we are primarily examining the power in terms of public sphere activities that are often dominated by elite men exercising power-over strategies that are assumed to be endemic to states.
6. This is also known as "Industry 4.0" (Furman & Seamans, 2019, p. 162).
7. Such lags have been evident in the cases of other GPTs throughout history, such as, for example, the integrated circuit (Brynjolfsson et al., forthcoming).
8. As we will discuss, trade agreements could specify international privacy standards and thus mitigate races to the bottom.

References

Acemoğlu, Daron. (2024). Harms of AI. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.). *The Oxford Handbook of AI Governance*. Oxford University Press.

Acemoğlu, Daron & Restrepo, Pascual. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *The American Economics Review 10* (6), 1488–1542.

Agrawal, Ajay, Gans, Joshua, & Goldfarb, Avi. (2018a). Economic Policy for Artificial Intelligence. National Bureau of Economic Research Working Paper Series.

Agrawal, Ajay, Gans, Joshua, & Goldfarb, Avi. (2018b). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business School Press.

Allen, Gregory C. (2017, December 21). Project Maven brings AI to the fight against ISIS. *Bulletin of the Atomic Scientists*. https://thebulletin.org/project-maven-brings-ai-fight-against-isis11374.

Allen, Gregory, & Chan, Taniel. (2017, July). "Artificial intelligence and national security." Study (Belfer Center for Science and International Affairs). https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

Allison, Graham T. (1969). Conceptual models and the Cuban missile crisis. *American Political Science Review 63* (3), 689–718.

Altmann, Jürgen, & Sauer, Frank. (2017). Autonomous weapon systems and strategic stability. *Survival 59* (5), 117–142.

Autor, David H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives 29* (3), 3–30.

Ayoub, Kareem, & Payne, Kenneth. (2016). Strategy in the age of artificial intelligence. *Journal of Strategic Studies 39* (5–6), 793–819.

Bachrach, Peter, & Baratz, Morton S. (1962). Two faces of power. *The American Political Science Review 56* (4), 947–952.

Biddle, Stephen. (2004). Military power: Explaining victory and defeat in modern battle. Princeton University Press.

Bloom, Nicholas, Sadun, Raffaella, & Van Reenen, John. (2012). Americans do IT better: US multinationals and the productivity miracle. *American Economic Review 102* (1), 167–201.

Bode, Ingvild. (2019). Norm-making and the Global South: Attempts to regulate lethal autonomous weapons systems. *Global Policy 10* (3), 359–364.

Bostrom, Nick. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Braumoeller, Bear. (2008). Systemic politics and the origins of great power conflict. *American Political Science Review 102* (1), 7793.

Brynjolfsson, Erik, Rock, Daniel, & Syverson, Chad. (2019). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (pp. 23–60). University of Chicago Press.

Brynjolfsson, Erik, & Hitt, Lorin M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives 14* (4), 23–48.

Carpenter, Charli R. (2005). "Women, children and other vulnerable groups": Gender, strategic frames and the protection of civilians as a transnational issue. *International Studies Quarterly 49* (2), 295–334.

Caruso, Catherine. (2017, March 2). Time to fold, humans: Poker-playing AI beats pros at Texas hold'em. *Scientific American*. https://www.scientificamerican.com/article/time-to-fold-humans-poker-playing-ai-beats-pros-at-texas-hold-rsquo-em/.

Claude, Inis. (1962). *Power and International Relations*. Random House.

Cockburn, Iain M., Stern, Scott, & Henderson, Rebecca. (2018). The impact of artificial intelligence on innovation. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press. (pp. 115–146)

Congressional Research Service (CRS). (2020, November 10). Intelligence and national security. https://crsreports.congress.gov R45178.

Crafts, Nicholas. (2004). Steam as a general purpose technology: A growth accounting perspective. *Economic Journal 114* (495), 338–351.

Crawford, Kate. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cummings, M. L., Roff, Heather M., Cukier, Kenneth, Parakilas, Jacob, & Bryce, Hannah. (2018, June). Artificial intelligence and international affairs: Disruption anticipated. Chatham House: The Royal Institute of International Affairs. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Daniels, Matthew, & Chang, Ben. (2021). National power after AI. Center for Security and Emerging Technology (CSET) Report. Georgetown University. https://cset.georgetown.edu/publication/national-power-after-ai/.

Dahl, Robert A. (1957). The concept of power. *Behavioral Science 2* (3): 201–215.

David, Paul A. (1990). The dynamo and the computer: An historical perspective on the productivity paradox. *American Economic Review 80* (2): 355–361.

Davis, Zachary. (2019). Artificial intelligence in the battlefield. *PRISM 8* (2), 118–121.

Ekelhof, Merel. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy 10* (3), 343–348.

Goldfarb, Avi, & Tucker, Catherine. (Forthcoming). "Inequality, privacy and digital market design." In Scott Kominers and Alex Teytelboym (Eds.), *More Equal by Design*. Oxford University Press.

Goldman, Emily O., & Leslie Eliason, eds. (2003). *The diffusion of military technology and ideas*. Stanford University Press.

European Union. (2019). EU guidelines on ethics in artificial intelligence: Context and implementation. European Parliamentary Research Service Briefing. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

Fearon, James. (1995). Rationalist explanations for war. *International Organization 49* (3), 379–414.

National Security Commission on Artificial Intelligence. (2021, March). Final Report. Washington, DC: National Security Commission on Artificial Intelligence.

Finnemore, Martha, & Sikkink, Kathryn. (1998). International norm dynamics and political change. *International Organization 52* (4), 887–917.

Frey, Carl B., & Osborne, Michael A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114*, 254–280.

Executive Office of the President. (2016, December). Artificial intelligence, automation, and the economy. Executive Office of the President, The White House. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

Furman, Jason, & Seamans, Robert. (2019). AI and the economy. *Innovation Policy and the Economy 19* (1), 161–191.

Garcia, Megan. (2017, June 3). How to keep your AI from turning into a racist monster. *Wired*.

Garfinkel, Ben, & Dafoe, Allan. (2019). How does the offense-defense balance scale? *Journal of Strategic Studies 42* (6), 749–753.

Geist, Edward. (2016). It's already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists 72* (5), 318–321.

Gehlhaus, Diana, Hodge, Ron, Koslosky, Luke, Goode, Kayla, & Rotner, Jonathan. (2021, September). The DOD's hidden artificial intelligence workforce. Center for Security and Emerging Technology.

Gill, Amandeep Singh. (2019). Artificial intelligence and international security: The long view. *Ethics & International Affairs 33* (2), 169–179.

Gilpin, Robert. (1981). *War and change in international politics*. Cambridge University Press.

Goolsbee, Austan. (2019). Public policy in an AI economy. In Ajay Agrawal, Joshua Gans and Avi Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 301–316). University of Chicago Press.

Grace, Katja, Salvatier, John, Dafoe, Allan, Zhang, Baobao, & Evans, Owain. (2018). Viewpoint: When will AI exceed human performance? Evidence from AI experts. *The Journal of Artificial Intelligence Research 62*, 729–754.

Graetz, Georg, & Michaels, Guy. (2018). Robots at work. *The Review of Economics and Statistics 100* (5), 753–768.

Halperin, Morton H., and Clapp, Priscilla. (2007). *Bureaucratic politics and foreign policy*. Brookings Institution Press.

Harris, Laurie A. (2017). Overview of artificial intelligence. U.S. Congressional Research Service Report, No. IF10608. https://crsreports.congress.gov/product/pdf/IF/IF10608.

Herz, John. (1950). Idealist internationalism and the security dilemma. *World Politics 2* (2), 157–180.

Horowitz, Michael C. (2010). *The diffusion of military power: Causes and consequences for international politics*. Princeton University Press.

Horowitz, Michael C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review 1* (3), 36–57.

Horowitz, Michael C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies 42* (6), 764–788.

Horowitz, Michael, Kahn, Lauren, & Mahoney, Casey. (2020). The future of military applications of artificial intelligence: A role for confidence-building measures? *Orbis 64* (4), 528–543.

Horowitz, Michael, & Scharre, Paul. (2021, January) AI and international stability: Risks and confidence-building measures. Center for a New American Security. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

Horowitz, Michael, Scharre, Paul, Allen, Gregory C., Frederick, Kara, Cho, Anthony, & Saravalle, Edoardo. (2018). Artificial intelligence and international security. Center for a New American Security. https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security.

Ikenberry, John. (2001). *After victory: Institutions, strategic restraint, and the rebuilding of order after major wars*. Princeton University Press.

Imbrie, Andrew, Fedasiuk, Ryan, Aiken, Catherine, Chhabra, Tarun, & Chahal, Husanjot. (2020). Agile alliances: How the United States and its allies can deliver a democratic way of AI. Center for Security and Emerging Technology (CSET) Report. Georgetown University. https://cset.georgetown.edu/publication/agile-alliances/.

Ipsos. (2019, January 22). Six in ten (61%) respondents across 26 countries oppose the use of lethal autonomous weapons systems. Press Release. https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons.

Jensen, Benjamin M., Whyte, Christopher, & Cuomo, Scott. (2020). Algorithms at war: The promise, peril, and limits of artificial intelligence. *International Studies Review 22* (3), 526–550.

Jervis, Robert. (1978). Cooperation under the security dilemma. *World Politics 30* (2), 167–214.

Jorgenson, Dale W., Ho, Mun S., & Stiroh, Kevin J. (2008). A retrospective look at the U.S. productivity growth resurgence. *Journal of Economic Perspectives 22* (1), 3–24.

Kania, Elsa B. (2017, November 28). *Battlefield singularity: Artificial intelligence, military revolution, and China's future military power*. Center for a New American Security.

Katzenstein, Peter J., ed. (1996). Introduction: Alternative perspectives on national security. In *The culture of national security: Norms and identity in world politics* (pp. 1–32). New York: Columbia University Press.

Keohane, Robert. (1984). *After hegemony: Cooperation and discord in the world political economy*. Princeton University Press.

Kennedy, Paul. (1987). *The rise and fall of the great powers*. Random House.

Konaev, Margarita, Chahal, Husanjot, Fedasiuk, Ryan, Huang, Tina, & Rahkovksy, Ilya. (2020, October). *U.S. military investments in autonomy and AI: Costs, benefits, and strategic effects*. Center for Security and Emerging Technology.

Lake, David. (2011). *Hierarchy in international relations*. Cornell University Press.

Lake, David, Martin, Lisa, & Risse, Thomas. (2021). Challenges to the liberal order: Reflections on international organization. *International Organization 75* (2), 1–33.

Lewis, Larry. (2018, March). Redefining human control: Lessons from the battlefield for autonomous weapons. Center for Naval Analyses. https://www.cna.org/CNA_files/PDF/DOP-2018-U-017258-Final.pdf.

Lin-Greenberg, Erik. (2020). Allies and artificial intelligence: Obstacles to operations and decision-making. *Texas National Security Review 3* (2), 57–76.

Maas, Matthijs. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy 40* (3), 285–311.

Madrigal, Alexis C. (2011, October 12). Siri: The perfect robot for our time. *The Atlantic*. https://www.theatlantic.com/technology/archive/2011/10/siri-the-perfect-robot-for-our-time/246516/.

March, James G., & Olsen, Johan P. (2011). The logic of appropriateness. In Robert Goodin (Ed.), *The Oxford Handbook of Political Science*. Oxford University Press. doi.org/10.1093/oxfordhb/9780199604456.013.0024, 478–497.

Market for Intelligence Conference. (2017, October 26). Toronto. https://www.youtube.com/watch?v=RxLIQj_BMhk.

Marshall, Andrew. (1966). *Problems estimating military power*. RAND.

McGrath, James R. (2016). Twenty-first century information warfare and the third offset strategy. *Joint Forces Quarterly 82* (3), 16–23.

Mearsheimer, John. (2019). Bound to fail: The rise and fall of the liberal international order. *International Security 43* (3), 7–50.

Melamud, Mordechai, Meertz, Paul, & Zartman, I. William, eds. (2014). Lessons from the CTBTO negotiation process. In *Banning the Bang Or the Bomb? Negotiating the Nuclear Test Ban Regime* (pp. 341–349). Cambridge University Press.

Meltzer, Joshua P. (2020, August 5). The court of justice of the European Union in Schrems II: The impact of GDPR on data flows and national security. Brookings Institution. https://www.brookings.edu/research/the-court-of-justice-of-the-european-union-in-schrems-ii-the-impact-of-gdpr-on-data-flows-and-national-security/.

Mercer, Jonathan. (1996). *Reputation and international politics*. Cornell University Press.

Metz, Cade. (2016, March 16). In two moves, AlphaGo and Lee Sedol redefined the future. *Wired*. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

Meyer, David. (2018, October 10). Amazon reportedly killed an AI recruitment system because it couldn't stop the tool from discriminating against women. *Fortune*. https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/.

Mittelsteadt, Matthew. (2021). AI verification: Mechanisms to ensure AI arms control compliance. Center for Security and Emerging Technology (CSET) Issue Brief. https://cset.georgetown.edu/publication/ai-verification/.

Müller, Vincent C., & Bostrom, Nick. (2014). Future progress in artificial intelligence: A survey of expert opinion. Fundamental Issues of Artificial Intelligence, 555–572. https://doi.org/10.1007/978-3-319-26485-1_33.

Ng, Andrew. (2017, April 28). Artificial Intelligence is the new electricity. *Medium*. https://medium.com/syncedreview/artificial-intelligence-is-the-new-electricity-andrew-ng-cc132ea6264.

Oliner, Stephen D., Sichel, Daniel E., & Stiroh, Kevin J. (2007). Explaining a productive decade. *Brookings Papers on Economic Activity 1*, 81–137.

O'Neil, Cathy. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Organski, A. F. K. (1958). *World Politics*. Alfred A. Knopf.

Poast, Paul. (2019). Beyond the "sinew of war": The political economy of security as a subfield. *Annual Review of Political Science 22*, 223–239.

Powell, Robert. (2006). War as a commitment problem. *International Organization 60* (1), 169–203.

Reeves, Shane R., Alcala, Ronald T. P., & McCarthy, Amy. (2020). Challenges in regulating lethal autonomous weapons under international law. *Southwestern Journal of International Law 27* (1), 101–118.

Risse, Thomas. (2000). "Let's argue!": Communicative action in world politics. *International Organization 54* (1), 1–39.

Roff, Heather. (2018). Advancing human security through artificial intelligence. In *Artificial Intelligence and International Affairs: Disruption Anticipated*. Chatham House Report. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Roff, Heather. (2019). The frame problem: The AI "arms race" isn't one. *Bulletin of the Atomic Scientists 75* (3), 95–98.

Rosenberg, Nathan. (1983). The effects of energy supply characteristics on technology and economic growth. In Sam Schurr, Sidney Sonenblum, and David O. Wood (Eds.), *Energy, productivity, and economic growth* (pp. 279–305). Oelgeschlager, Gunn, and Hain.

Rosenberg, Nathan. (1998). The role of electricity in industrial development. The Energy Journal 19(2), 7–24.

Rosendorf, Ondřej. (2021). Predictors of support for a ban on killer robots: Preventive arms control as an anticipatory response to military innovation. *Contemporary Security Policy 42* (1), 30–52.

Rosert, Elvira, & Sauer, Frank. (2021). How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy 42* (1), 4–29.

Ruggie, John. (1982). International regimes, transactions, and change: Embedded liberalism in the postwar economic order. *International Organization 36* (2), 379–415.

Runyan, Anne Sisson, & Peterson, V. Spike. (2014). *Global gender issues in the new millennium*. 4th ed. Westview.

Ryan, Mick. (2017, December 11). Building a future: Integrated human–machine military organization. *Strategy Bridge*. https://thestrategybridge.org/the-bridge/2017/12/11/building-a-future-integrated-human-machine-military-organization.

Scharre, Paul. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton.

Scharre, Paul. (2021). Debunking the AI arms race theory. *Texas National Security Review 4* (3), 121–132.

Scharre, Paul, & Horowitz, Michael C. (2015). An introduction to autonomy in weapon systems. Center for a New American Security Working Paper (February). https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.

Scharre, Paul. (2016, February). *Autonomous weapons and operational risk*. Center for a New American Security.

Schurre, Sam H, Burwell, Calvin C, Devine, Warren S, & Sonenblum, Sidney. (1990). *Electricity in the American economy: Agent of technological progress*. Greenwood Press.

Simon, Herbert A. (1995). Artificial intelligence: an empirical science. *Artificial Intelligence 77* (1), 95–127.

Snyder, Glenn. (1997). *Alliance politics*. Cornell University Press.

Snyder, Jack. (1984). Civil-military relations and the cult of the offensive, 1914 and 1984. *International Security 9* (1), 108–146.

Strange, Susan. (1995). The defective state. *Daedalus 123* (2), 55–74.

Tambe, Prasanna, Cappelli, Peter, & Yakubovich, Valery. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review 61* (4), 15–42.

Talmadge, Caitlin. (2019). Emerging technology and intra-war escalation risks: Evidence from the Cold War, implications for today. *Security Studies 42* (6), 864–887.

Tannenwald, Nina. (1999). The nuclear taboo: The United States and the normative basis of nuclear non-use. *International Organization 53* (3), 433–468.

Tarraf, Danielle C., Shelton, William, Parker, Edward, Alkire, Brien, Gehlhaus, Diana, Grana, Justin, Levedahl, Alexis, Léveillé, Jasmin, Mondschein, Jared, Ryseff, James, Wyne, Ali, Elinoff, Daniel, Geist, Edward, Harris, Benjamin N., Hui, Eri, Kenney, Cedric, Newberry, Sydne J., Sachs, Chandler, Schirmer, Peter, Schlang, Danielle... Warren, Kristin (2019). "The Department of Defense posture for artificial intelligence. RAND Corporation. https://www.rand.org/pubs/research_briefs/RB10145.html.

Tucker, Catherine. (2019). "Privacy, Algorithms, and artificial Intelligence." In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (pp. 423–38. University of Chicago Press.

Tucker, Patrick. (2015, February 12). The marines are building robotic war balls. *Defense One*. http://www.defenseone.com/technology/2015/02/marines-are-building-robotic-war-balls/105258/?oref=search_marines%20war%20balls.

U.S. Department of Defense. (2021, September 16). JAIC facilitates first-ever International AI dialogue for defense. Press Release. https://www.ai.mil/news_09_16_20-jaic_facilitates_first-ever_international_ai_dialogue_for_defense_.html.

Walt, Stephen. (1987). *The origins of alliances*. Cornell University Press.

Walter, Stefanie. (2021). The backlash against globalization. *Annual Review of Political Science 24*, 1.

Waltz, Kenneth. (1979). *Theory of international politics*. McGraw-Hill.

Wilner, Alex, & Babb, Casey. (2020). New technologies and deterrence: Artificial intelligence and adversarial behaviour. In Frans Osinga and Tim Sweijs (Eds.), *NL ARMS Netherlands annual review of military studies 2020: Deterrence in the 21st century—Insights from theory and practice* (pp. 401–417). Netherlands Defense Academy.

Wilson, James Q. (1989). *Bureaucracy: What government agencies do and why they do it*. Basic Books.

Wingfield, Nick. (2018, January 21). Inside Amazon Go, a store of the future. *The New York Times*. https://www.nytimes.com/2018/01/21/technology/inside-amazon-go-a-store-of-the-future.html.

CHAPTER 46

THE GHOST OF AI GOVERNANCE PAST, PRESENT, AND FUTURE

AI Governance in the European Union

CHARLOTTE STIX

Introduction

This chapter will provide a simplified overview of the past, present, and future of AI governance in the EU. It will give the reader a solid background understanding of how the EU reached the current status as global leader in the regulation of AI, how all the different pieces are interconnected, and where the EU might go next. First, I will discuss a select number of EU policy efforts of the past years and illustrate how they built on each other. It is argued that, by virtue of spearheading "trustworthy AI," the EU has occupied a position where it has been able to shape the discourse on global AI governance discourse early on. Then, I will introduce the history of the EU's AI regulation, proposed in April 2021, highlighting its connection to previous policy efforts, and its roots in adjacent measures to strengthen the EU's technological ecosystem. Finally, the discussion will turn to the future, considering and exploring a number of AI governance areas that are prime candidates to become crucial for AI governance in the EU in the coming decade.

The Past

Taking stock: The roads towards the EU's AI governance

This section will highlight the most relevant and recent EU policy developments with regard to AI, and will illustrate how they contributed to shaping both the broader narrative for the EU and how they set the cornerstones for AI governance in the EU. It will be suggested that the AI Act (European Commission, 2021c), the European Commission proposal for a

horizontal AI regulation, and the accompanying policy measures in the EU form a coherent and strategically aligned link in a chain of policies which were initiated many years ago. To that end, some of these policy documents will be revisited in the chapter, presenting the different elements that form the bigger picture of current AI governance in the EU.

While going through the formative policy developments, it is worth underlining that the European Commission has invested in and funded AI and AI-related research and innovation projects for much longer than they have been focusing on the governance of AI, notably, under Horizon 2020 and before. With that in mind, the following discussion will set out the main elements that led to the EU to push for ethical governance of AI and to make it their guiding principle for accompanying policy measures.

The roads that led us here

The earliest key policy document is the resolution by the European Parliament with "Recommendations to the Commission on Civil Law Rules on Robotics" in 2017 (henceforth: "Civil Law Rules on Robotics"; European Parliament, 2017). Although not yet referring to AI directly in the title, the resolution laid one of the first cornerstones for the succeeding process from the European Parliament's side, by suggesting that the EU's legal framework should be updated and complemented by ethical principles on the topic, that the environmental impact of AI and robotics should be kept low, and that the societal and economic impacts of future systems deserve heightened attention.[1]

Shortly thereafter, the European Economic and Social Committee (EESC) presented their "Opinion on AI" (European Economic and Social Committee, 2017). The "Opinion on AI" (European Economic and Social Committee, 2017) discusses the need to verify, validate, and monitor AI and AI-based systems, advocates for an overarching "human-in-command" approach, and leans into the necessity for ethical, societal, and safety considerations. Accordingly, recommendations cover the development of a code for ethics, a ban on Lethal Autonomous Weapons Systems, and the development of suitable standardization systems for AI.

On evaluation, we can already see that while they are among the earliest EU policy documents on the topic, the European Parliament's (EP) "Civil Law Rules on Robotics" (European Parliament, 2017) and the EESC's "Opinion on AI" (European Economic and Social Committee, 2017) have some overlaps. These include a demand for inclusion of the ethical dimension in the discussion and an acknowledgement of the societal impact, alongside proposals for recourse.

The shift towards EU AI governance as a topic largely independent of robotics from the perspective of EU institutions was further solidified through the European Group on Ethics in Science and New Technologies (EGE) Statement on "Artificial Intelligence, Robotics and Autonomous Systems" (henceforth: "Statement"; European Group on Ethics in Science and New Technologies, 2018). An independent advisory body to the European Commission, the EGE advises it on the intersection of science and emerging technologies with ethical, societal, and fundamental rights issues. In their "Statement" (European Group on Ethics in Science and New Technologies, 2018), they echoed the need to establish an overarching framework on AI in the EU with an ethical dimension. The goal would be to tackle the ethical, legal, and societal governance issues, ensuring that AI is created with "humans in mind" (European Group on Ethics in Science and New Technologies, 2018).

Therefore, in a sense building on the resolution in "Civil Law Rules on Robotics" (European Parliament, 2017) and "Opinion on AI" (European Economic and Social Committee, 2017), they proposed the development of several ethical principles for AI based on fundamental European values. This proposal was both quintessentially European, outlining the importance of fundamental rights and values, and well timed to fit within the broader international landscape, where principles for AI were starting to see their advent (Fjeld et al., 2020; Hagendorff, 2019; Schiff et al., 2020; Zeng et al., 2018).

These three documents together could be seen as the first heralds of where the EU is now: regulating AI with a focus on human-centricity and ethics. They demonstrated what aspects of AI were considered important by the EU's legislative body, the EU's civil society organization body, and the main EU group on ethics at that point. This is a convergence point where demonstrable attention has begun from a legislative angle, a societal angle and an ethical angle. As we will see, this interplay has since continued, and even been strengthened. At this point in time it became strikingly evident that the EU policy space was alert to the challenges posed by AI, as much as to the opportunities AI could hold, and the necessity for a more methodological approach became pressing.

What followed were major leaps. First, the European Commission presented the "Digital Day Declaration on Cooperation on AI" (European Commission, 2018c) in April 2018. Second, they responded to the call from the European Council to "put forward a human-centric approach to AI"[2] by presenting their AI strategy in the communication entitled, "Artificial Intelligence for Europe" (European Commission, 2018a), which was released in the same month. I propose that these documents foreshadow current EU AI governance mechanisms. The "Digital Day Declaration on Cooperation on AI" (henceforth: "Declaration"; European Commission, 2018c) anticipated the "Coordinated Plan on AI" (European Commission, 2018b), whereas the "Communication on AI for Europe" (European Commission, 2018a), in some sense preceded the proposal for a regulatory framework, the AI Act (European Commission, 2021c).

So, how did these documents set out the European Commission approach to AI governance?

In the "Declaration" (European Commission, 2018c), signed by all 28 Member States in 2018 (at that time including the United Kingdom), as well as Norway, the countries agreed to engage in close dialogue with the European Commission on the topic of AI and coordinate their actions. I propose that this is the first instance internationally where a significant number of countries agreed to coordinate on AI governance.[3] International, later-stage efforts to coordinate among multiple countries such as the Global Partnership on AI (GPAI) or the OECD included the EU by way of representation through the European Commission, as well as a subset of member states. As of this writing, there is no other international AI governance effort that envisages the same level of coordination, alignment of approach, and pooling of resources as the one started with the "Declaration"[4] and signed off by the member states.[5]

We have seen that the "Opinion on AI" (European Economic and Social Committee, 2017) called for a "human in command" approach and that the EGE "Statement" (European Group on Ethics in Science and New Technologies, 2018) called for AI to be made with "humans in mind." This notion of human-centric AI is revisited in the "Declaration" (European Commission, 2018c), committing signatories to ensure that "humans remain at the center of AI development," and to prevent the "harmful creation and use of AI applications."

Echoing topics outlined in the resolution on "Civil Law Rules of Robotics" (European Parliament, 2017) and the "Opinion on AI" (European Economic and Social Committee, 2017), the "Declaration" (European Commission, 2018c) focuses on the development of a collaborative framework to coordinate on relevant areas such as sustainability, labor market, funding, and ethics. Alongside the necessary mitigation of ethical risks, it also touches upon legal and socio-economic risks of AI. All of this is underlined with the recognition that the existing ecosystem needs to be boosted in order for the EU to remain competitive and agile for future challenges.

The "Communication on AI for Europe" (henceforward: "AI Strategy"; European Commission, 2018a) picked up on this and presented the EU's three-pronged strategy for AI taking into account the ecosystem, civil society as well as ethics and regulation. In short, it recommended to "(1) boost Europe's technological and industrial capacity; (2) to prepare Europe for the socio-economic changes associated with AI; and (3) to ensure that Europe has an appropriate ethical and legal framework to deal with AI development and deployment" (European Commission, 2018a).

The "AI Strategy" (European Commission, 2018a) also outlined many ambitions that are currently relevant, such as the development of regulatory sandboxes (eventually key for a horizontal EU AI regulation) and a commitment to support centers for data sharing (for example, the EU Data Hubs).

In order to tackle the third pillar of its "AI Strategy" (European Commission, 2018a), the European Commission set up an independent High-Level Expert Group on Artificial Intelligence (AI HLEG) tasked with, amongst other deliverables, the development of Ethics Guidelines.[6]

Moreover, the "AI Strategy" (European Commission, 2018a) also served to present, for the first time, a clear "European way for AI" vis-à-vis the international stage. It clearly outlined the role that the EU envisages for itself when it comes to AI governance. It stated that:

> The EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation and respects Union's values and fundamental rights as well as ethical principles such as accountability and transparency. The EU is also well placed to lead this debate on the global stage. This is how the EU can make a difference—and be the champion of an approach to AI that benefits people and society as a whole.
>
> (European Commission, 2018a)

The EU clearly positioned itself as an actor on the international stage who will put ethical considerations and fundamental rights at the core of AI governance.

Finally, we end this section by pulling some of the threads together while leaving space to next investigate the regulatory efforts. Many of the efforts highlighted culminate or find resonance in the European Commission's "Coordinated Plan on the Development and Use of Artificial Intelligence Made in Europe" (henceforth: "Coordinated Plan"; European Commission, 2018b) published in late 2018.

The "Coordinated Plan" (European Commission, 2018b) picks up where the "Declaration" (European Commission, 2018c) left off. It too was agreed on by all Member States, as well as Norway and Switzerland, and is to be updated on a rolling basis. It echoes plans mapped out in the "AI Strategy" (European Commission, 2018a), namely that a European approach to AI should be built upon ethical and societal values derived from the Charter of Fundamental Rights. Moreover, going above and beyond the perspective of previous policy

documents, it puts a strong emphasis on what should become the European north star for AI, by highlighting what it perceives to be interconnected concepts of "trusted AI" and "human-centric AI" (European Commission, 2018b).

The "Coordinated Plan" (European Commission, 2018b) paints a picture of how the Member States can coordinate their AI strategies, define a common vision, and encourage synergies between ongoing efforts in the Member States—with an eye to increasing the EU's global competitiveness and to counteracting fragmentation and competition between like-minded actors. The preliminary framework for coordination homes in on a couple of focus areas such as on commonly shared societal challenges, increased diffusion of AI, support to AI excellence, data availability, and a regulatory framework. The last aspect is described as a "seamless regulatory environment" in other parts of the text, and we will see in the following discussion what the EU has developed on that front. The "Coordinated Plan on AI" (European Commission, 2018b) also contained a commitment on the EU's side to invest EUR 20bn into AI by 2020, and scale up towards yearly investments of that sum from then until 2027.[7]

In an adjacent stream, in early 2019, the European Parliament's Committee on Industry, Research and Energy (ITRE) had their report on "A Comprehensive European Industrial Policy on Artificial Intelligence and Robotics" (European Parliament, 2019) adopted, which articulated a clear need for a "robust legal and ethical framework for AI," and which, amongst other things, called for ethical principles that are in compliance with relevant EU and national law. Along these lines it also welcomes the work of the AI HLEG. Furthermore, it proposes aspects related to personal data and privacy, consumer protection, transparency, explainability, and bias. In tandem with the EU's approach (as mapped so far), the industrial policy stresses the importance of human-centric technology and to encourage ethical values with regards to AI development and deployment. Indeed, this may set the EU apart and propel it to take lead on an international stage.

Combining the various policy efforts and taking a bird's-eye view, clear directions are emerging that concern the role the EU has set for itself generally and on the international stage. Ethical concerns, fundamental rights, and values play a vital role in the EU's AI governance future. To account for this, this chapter will next focus on the development of ethical principles for AI in the EU—and their subsequent impact. This should be seen as an adjacent stream of work that resulted out of the landscape built by the policy initiatives outlined here which became a policy effort in its own right, eventually feeding back into the current landscape.

Coining "trustworthy AI"

This section explores how the EU came to adopt and pioneer the term "trustworthy AI" from its ethical investigations, and what this shift marked.

Subsequent to its ambition to develop an appropriate ethical and legal framework for AI, the European Commission set out to establish two groups to support the AI strategy described in its Communication on AI: the High-Level Expert Group on AI (AI HLEG) and the European AI Alliance. The latter was set up as an accessible online multi-stakeholder platform with the goal of contributing to the work of the European Commission and the AI HLEG.

The AI HLEG, on the other hand, was set up as an independent expert group by the European Commission, populated through a selection process (Stix, 2021). The AI HLEG

was tasked with the primary goal of developing ethics guidelines. The result of their work, especially the "Ethics Guidelines for Trustworthy AI" (AI HLEG, 2019) and the "Assessment List for Trustworthy AI" (henceforth: "Assessment List"; AI HLEG, 2020), were core to the recent model of AI governance in the EU. To that end, the focal point for the following discussion will be on the "Ethics Guidelines for Trustworthy AI" (A HLEG, 2019) and the associated "Assessment List" (AI HLEG, 2020), and how the conceptualization of ethical AI contained in them contributed to Europe's vision of "trustworthy AI."

Tasked with the development of ethics guidelines, the AI HLEG, composed of 52 experts representing various sectors and types of expertise, underwent a comprehensive process to take a unique step towards a framework for the ethical governance of AI. Although the work was conducted internally, the AI HLEG did share their progress in meetings open to institutional observers and solicited feedback on their first draft version of the Ethics Guidelines six months into the process via the AI Alliance.[8] Following the implementation of this public feedback, the AI HLEG presented their final "Ethics Guidelines for Trustworthy AI" (henceforth: "Ethics Guidelines"; AI HLEG, 2019) in April 2019.

The "Ethics Guidelines" (AI HLEG, 2019) constituted the first document that proposed a clear conceptual understanding and framing of what type of AI should be encouraged within the EU. While the document is strongly anchored in EU values and fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union, the core concept is that of "trustworthy AI." In this reading, "Trustworthy AI" is to fulfill three conditions: (1) it should be lawful, (2) it should be ethical, and (3) it should be robust (both from a technical and social perspective).

While the "lawful" aspect is left to existing regulation and future regulatory efforts, the document proceeds to outline the other components. In particular, our focus will be on the ethical component. The AI HLEG distilled a number of core values, which informed four principles. These are; *Respect for Human Autonomy*, *Prevention of Harm*, *Fairness*, and *Explicability*. From these four ethical principles, they derived their seven key requirements to achieve "trustworthy AI" and to operationalize these four identified principles.

The seven key requirements covered:

- *Human Agency and Oversight*, which relates to the principle of *Human Autonomy* and requires that AI systems allow for human oversight, support the user's agency, and foster fundamental rights.
- *Technical Robustness and Safety*, which relates to the principle of *Prevention of Harm*. It addresses concerns such as resilience to attack (e.g., through data poisoning or model leakage), the need for suitable fallback plans, reliability, and reproducibility.
- *Privacy and Data Governance*, which links to the principle of *Prevention of Harm*. It tackles the initial stages of data collection (e.g., regarding the quality and integrity of the data), as much as the need for data protocols to govern data access, and overall privacy measures throughout the AI life cycle.
- *Transparency*, which links to the principle of *Explicability*. This means, among other things, that traceability should be ensured and that capabilities and intentions (both from a technical POV and from an industry perspective) should be clearly communicated.

- *Diversity, Non-Discrimination, and Fairness*, which links to the principle of *Fairness*. It states that all affected stakeholders throughout the AI life cycle need to be taken into consideration and duly involved (e.g., ensuring equal access and equal treatment).
- *Societal and Environmental Well-Being*, which relates to both the principle of *Fairness* and the principle of *Prevention of Harm*. It relates to the broadest range of stakeholders, the environment and the wider society. Considerations include, for example, AI's social impact and the sustainability of the current AI supply chain.
- *Accountability*, which is the last key requirement and ties all the previous requirements together. It is informed by the principle of *Fairness*. It focuses on redress mechanisms, trade-offs between principles, and the need to have adequate mechanisms in place to report potential negative impacts.

At this point, the threads started with the report on "Civil Law Rules on Robotics" (European Parliament, 2017), the "Opinion on AI" (European Economic and Social Committee, 2017), and the "AI Strategy" (European Commission, 2018a) have come to reach a fuller picture: the EU's ambition to create an ethical approach towards the development and deployment of AI and an appropriate ethical framework has become a reality.

In order to operationalize these key requirements further, the "Ethics Guidelines" (AI HLEG, 2019) also contained a draft "Assessment List" which was piloted and revised in the second year of the group's mandate.[9] The final "Assessment List for Trustworthy AI: for Self-Assessment" (AI HLEG, 2020) is the first tool in the EU that took "trustworthy AI" into account throughout an AI system's lifecycle and outlined how an assessment of that could take shape. It was also among the earliest serious attempts to translate ethical principles for AI into actionable measures for all stakeholders involved throughout the AI lifecycle, be that researchers, industry, government, or civil society.

The concept of "trustworthy AI," in the way that the AI HLEG defined it, became a cornerstone for AI governance in the EU. Building on its previous emphasis on "human-centric AI" (as we have seen in the European Council's call to the European Commission,[10] the "Declaration" [European Commission, 2018c], and the "Coordinated Plan" [European Commission, 2018b]), the European Commission adopted this conceptual approach and expanded upon it in the Communication on "Building Trust in Human-Centric AI" (European Commission, 2019). In that communication, the European Commission supports the key requirements and the concept of trustworthy AI, stating that:

> Only if AI is developed and used in a way that respects widely-shared ethical values, it can [sic] be considered trustworthy.

This can be understood—and as we will see, is reflected in subsequent governance documents and decisions—as the European Commission embracing the concept of "trustworthy AI" as a core component of its strategic vision.[11]

Equally, it doubles down on the reputation of the European Union as a region that produces "safe and high-quality products" (European Commission, 2019). To consolidate its place as a leader on "trustworthy AI" on an international stage, the Communication on "Building Trust in Human-Centric AI" (European Commission, 2019) furthermore launched a consensus building an "International Alliance for a human-centric approach

to AI."[12] Its goal is to share the EU's vision and ambitions with like-minded international partners.

Beyond that, "Building Trust in Human-Centric AI" (European Commission, 2019) further expands on elements of documents such as the "Coordinated Plan" (European Commission, 2018b). It reiterates the core foci to boost the ecosystem, such as an increase in joint ventures, pooling of data and other building blocks for AI, as well as the strengthening of synergies across Member States. It proposed to launch a set of networks of AI research excellence centers under the Horizon 2020 research and innovation framework program, to set up networks of AI-focused Digital Innovation Hubs (DIHs) and to develop and implement a model for data sharing and common data spaces amongst Member States and other stakeholders. These suggestions directly shape the current state of affairs as we will see next.

The Present

The third way: The EU's AI north star

We have now reviewed the recent historical backdrop of the current EU's regulatory strategy, discussing both its roots and predecessors. This brings us to today. The EU is looking to develop an attractive alternative to U.S. and Chinese approaches to AI governance. In order to gain a bird's-eye view of that third way, I will highlight a select number of important and current developments, illustrating how they contribute to the EU's direction.

On February 19, 2020, the European Commission published a comprehensive package consisting of: the "European Strategy for Data" (European Commission, 2020a), the report on "Safety Liability and Implications of AI, the Internet of Things and Robotics" (European Commission, 2020f), and the "White Paper on Artificial Intelligence: A European Approach to Excellence and Trust" (henceforth: "White Paper on AI"; European Commission, 2020d). Due to the limited scope of this chapter, we will focus on the "White Paper on AI" (European Commission, 2020d).

The "White Paper on AI" (European Commission, 2020d) followed European Commission president von der Leyen's promise in her political agenda to put forward "legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence."[13] It solidified the commitment to human-centric and "trustworthy AI," adding the first step towards a future legislative framework built on the concept of "trustworthy AI" to the new EU AI governance portfolio. The "White Paper on AI" (European Commission, 2020d) is divided into two main sections, one on an Ecosystem of Trust, focusing on the first proposal for a regulatory framework, and one on an Ecosystem of Excellence, focusing on supporting the European AI ecosystem. Both of these closely match ambitions outlined in previous policy documents as previously discussed, such those in Europe's AI strategy (European Commission, 2018a), and fit in with other recent governance efforts. I will therefore use the "White Paper on AI's" (European Commission, 2020d) duality of policy and infrastructure to highlight how far the EU AI policy has come in each area since this chapter's introductory Section and how they build on one another to make the EU a hub for "trustworthy AI."

On April 21, 2021, the European Commission published its package on a European approach for AI containing: a "Communication on Fostering a European Approach to Artificial Intelligence" (European Commission, 2021a); a "Coordinated Plan on AI: 2021 Review" (European Commission, 2021b); and, the highly anticipated proposal for a "Regulation on a European Approach for Artificial Intelligence (AI Act)" (European Commission, 2021c). The "Coordinated Plan on AI: 2021 Review" (European Commission, 2021b) builds on the "Coordinated Plan" (European Commission, 2018b) and dramatically expands its scope and ambitions, and the "Regulation on a European Approach for Artificial Intelligence (AI Act)" (European Commission, 2021c) builds on the "White Paper on AI" (European Commission, 2020d) and subsequent impact assessments conducted by the European Commission.[14] We will now see how all of this shapes up to form the context and the ecosystem for future AI governance in the EU.

Trust and the EU's AI governance

The chapter in the "White Paper on AI" (European Commission, 2020d) dedicated to the Ecosystem of Trust outlined the European Commission's policy proposals for a potential regulation prior to the final publication of the proposal for a "Regulation on a European Approach for Artificial Intelligence (AI Act)" (European Commission, 2021c) on April 21, 2021. The chapter was strongly inspired by the conceptual idea of "trustworthy AI" and heavily referenced the work of the AI HLEG.

The core proposal suggested that in an envisioned horizontal legislation mandatory legal requirements should apply to high-risk cases of AI only. These high-risk AI cases were defined by the following cumulative criteria: if the sector itself is high risk (e.g., healthcare, transport) and if the intended use involves high risk (e.g., injury, death, significant material/immaterial damage).

The mandatory legal requirements largely reflect the "Ethics Guidelines" (AI HLEG, 2019) seven key requirements and are composed of the following: a requirement for adequate training data, a requirement for data and record keeping, a requirement for the provision of information, a requirement for robustness and accuracy, and a requirement on human oversight. The final requirement was specifically laid out for the case of remote biometric identification. High-risk AI systems would be subject to conformity assessment (e.g., testing, inspection, and certification) accounting for these requirements before they would be able to enter the EU market.

The "White Paper on AI" (European Commission, 2020d) also outlined strategies for non-high-risk AI systems. It was suggested that these could partake in a voluntary labeling scheme which could build upon or implement the "Assessment List" (AI HLEG, 2020). We see that the building blocks and vision previously discussed are starting to take considerable shape building the EU's AI governance future.

Soon after, the legal affairs committee of the European Parliament adopted several aligned reports. These tackled an ethical framework for AI, civil liability claims against operators of AI systems, and the protection of intellectual property rights with regards to AI.[15] It is noteworthy that the first report's guiding principles strongly resembled those of the "Ethics Guidelines." We can infer that the vision of EU AI governance is coherent across the EU institutions, which is important[16] as we move to the most recent and high-profile

policy development on the European Commission's side: the "Regulation on a European Approach for Artificial Intelligence (AI Act)" (European Commission, 2021c).

Following the publication of the "White Paper on AI" (European Commission, 2020d), the European Commission conducted impact assessments and opened a stakeholder consultation to receive feedback on the "White Paper on AI" (European Commission, 2020d). This feedback shaped the subsequent proposal for a regulation.

The proposed "Regulation on a European Approach for Artificial Intelligence (AI Act)" (henceforth: "AI Act"; European Commission, 2021c) introduces the European Union's legislation for AI, specifically high-risk AI systems. It is a risk-based regulation which covers stand-alone AI systems that are considered high-risk which are elaborated on in Annex III to the "AI Act" and cover use cases such as in law enforcement for individual risk assessment, education and vocational training for determining access to educational or training institutions, or specific cases of access to essential public and private services and benefits. In short, Annex III lists a number of areas and specific use cases in those areas where stand-alone AI systems will automatically be considered high risk due to their potentially adverse impact on health, safety, or the fundamental rights of persons or groups. The other case of high-risk AI systems are those that are not stand-alone AI systems but those that are safety components of products or systems, or those that are products or systems.

Both of these types of high-risk AI systems need to comply with a number of requirements the "AI Act" (European Commission, 2021c) lays down in Title III Chapter II, although the manner in which that compliance is achieved, documented, and assessed (conformity assessment) is different between stand-alone and integrated high-risk AI systems. In the case of high-risk AI systems that are safety components of products or systems, or are themselves products or systems, the harmonized "AI Act" (European Commission, 2021c) adjusts to fit with the existing sectoral procedures, rules, and regulations.

The scope of this "AI Act" (European Commission, 2021c) encompasses a range of actors: providers that place their AI system on the EU market, users of AI systems in the EU (except those that use it in a personal, non-professional activity), and providers and users of AI systems that are not based in the EU but where the output of their AI system is used in the EU.

All high-risk AI systems need to fulfill the requirements set out in the "AI Act" (European Commission, 2021c) Title III Chapter II. These requirements closely match those that were previously proposed in the "White Paper on AI" (European Commission, 2020d) and, as we have seen, are therefore closely connected to the requirements within the "Ethics Guidelines" (AI HLEG, 2019). The requirements listed in the "AI Act" are: *Data and Data Governance*; *Technical Documentation*; *Record Keeping*; *Transparency and Provision of Information to Users*; *Human Oversight*; and *Accuracy, Robustness, and Cybersecurity*.

Whilst they are not described in this format, I would like to propose that the requirements can be thought of in two categories: those that are procedural and those that are informative. *Data and Data Governance*; *Human Oversight*; and *Accuracy, Robustness, and Cybersecurity* are procedural. They concern themselves with the workings of the algorithm throughout its lifecycle and how these can be affected in a positive manner to avoid negative impacts. By contrast, *Technical Documentation*, *Record Keeping*, and *Transparency and Provision of Information to Users* can be considered as informational requirements. They track procedural information, check it, and monitor it throughout the AI system's life cycle.

Those actors that are responsible for ensuring that a high-risk AI system complies with the "AI Act" must fulfill certain conditions on top of adherence to the requirements mentioned previously. In short, they must first build their AI system in accordance with the requirements from Title III Chapter II, then they have to undertake an internal conformity assessment of the AI system which encompasses paper trails and documentation generated in the first step and developed as a framework for the AI system throughout its functioning. That will entail a Quality Management System which ensures compliance with the "AI Act" (European Commission, 2021c), a Risk Management system which acts as a continuous iterative process throughout the AI system's lifecycle, and Technical Documentation which covers elements such as detailed descriptions, pre-determined changes of the AI system and the performance, as well as monitoring, functioning, and control of the AI system. Third, the provider or relevant other actor needs to establish a post-market monitoring system for the AI system once it has been put on the market or placed into service. This post-market monitoring system will collect logs produced by the AI system, act as a supervisor to the AI system, and report serious incidents if they occur. Finally, before the AI system can be put on the market or placed into service it must be registered in the EU database, and EU Declaration of Conformity must be filled out to describe its adherence to the "AI Act" (European Commission, 2021c), and it should be affixed with a CE marking to indicate that it has passed its conformity assessment.

In addition to requirements and procedures for high-risk AI systems, the "AI Act" (European Commission, 2021c) also lays down a number of AI systems that are prohibited for use in the EU under certain conditions. Without enumerating them in detail, these prohibited AI systems cover those that deploy subliminal techniques that could cause harm, those that exploit vulnerabilities in a manner that would cause harm, and those that public authorities could use to evaluate the trustworthiness of an individual, leading to unfavorable treatments in different contexts or treatment that is disproportionate. Moreover, it includes "real-time" biometric identification systems if they are used in publicly accessible spaces and for the purpose of law enforcement. However, noteworthy exceptions to the latter are cases where there is a targeted search for potential victims of crime, where it is in the public interest to prevent specific, substantial, and imminent threats and to detect certain perpetrators.

Akin to the proposals in the "White Paper on AI" (European Commission, 2020d), the "AI Act" (European Commission, 2021c) also briefly concerns itself with voluntary Codes of Conduct for non-high risk AI systems, with the intention to foster "trustworthy AI" and, therefore, compliance to the "AI Act" within the broader ecosystem. I will next discuss how the corresponding environment within the EU is boosted in order to establish the overarching framework that these policy and legislative ambitions would fit in with. Finally, we will discuss in more detail the specific elements of the ecosystem that are likely to become crucial to the EU's success in AI governance in the near future.

Strengthening the AI ecosystem

In recent years, it has become clear that the EU does not solely wish to rely on their regulatory expansionism, exporting norms and legislative approaches towards "trustworthy AI" on an international stage. Acknowledging that its ecosystem has, at times, difficulty

competing with tech giants developing or established outside of the EU, it is increasingly moving towards Digital Sovereignty. This encompasses the broader AI landscape in the EU. To have a truly comprehensive and integrated approach towards AI governance, ethical, policy, and regulatory efforts must be boosted in tandem with the existing and foreseen landscape. In short, an increase in relevant EU infrastructure for AI development, deployment, and use equals an increase in ownership of the technology, an increase in the ability to shape it directly through norms for trustworthy AI (through soft and hard law), and a decrease in reliance on outside actors. Keeping this in mind, the following paragraphs will sketch how the EU is building this infrastructure and what benefit this may yield, starting with the chapter in the "White Paper on AI" on an Ecosystem of Excellence and expanding it further with efforts outlined in the "Coordinated Plan on AI: 2021 Review" (henceforth: "Coordinated Plan: 2021 Review"; European Commission, 2021b) and adjacent initiatives.

Many of the areas below directly link back to aspects mentioned earlier in this chapter, such as in the "Declaration" (European Commission, 2018c), the "Coordinated Plan" (European Commission, 2018b), and in the "AI Strategy" (European Commission, 2018a), which all outlined the need to boost the ecosystem, to combine resources, and to increase technical capabilities to ensure the EU's leadership in human-centric and "trustworthy AI."

The "White Paper on AI" (European Commission, 2020d) includes a chapter on an Ecosystem of Excellence, which concerns itself with technical infrastructure as well as with ecosystem building. On the latter, it especially focuses on building new infrastructures. On the more research-oriented side, it proposed work on establishing a lighthouse center of research, innovation, and expertise to combat a seemingly fragmented AI research community in the EU. Looking at industry, small- and medium-sized enterprises (SME), and start-ups, it calls for the development of testing and experimentation facilities (TEFs), building out capacity via the Digital Innovation Hubs (as mentioned in the "Coordinated Plan"; European Commission, 2018b), a recently funded AI-on-Demand platform,[17] and engagement of key stakeholders through a Public–Private Partnership on AI, data, and robotics in the context of Horizon Europe.[18] It also touches on turning the "Assessment List" (AI HLEG, 2020) into an indicative curriculum for those developing AI, and an ambition to keep talent and attract talent to the EU through new education networks under the Digital Europe program.[19]

In accordance with the "European Data Strategy" (European Commission, 2020a), the "White Paper on AI" also puts an emphasis on the role of data in AI development ("compliance of data with the FAIR principles will contribute to build trust and ensure re-usability of data" European Commission, 2020d), as well as the importance of computing infrastructure. This will be explored in more detail at the end of this section.

Every single one of these proposed efforts ties in with the "Coordinated Plan: 2021 Review" (European Commission, 2021b) and demonstrates the EU's efforts to build an infrastructure that can match its ambition on the governance side to promote the development of human-centric, sustainable, inclusive, and trustworthy AI. While the "Coordinated Plan" (European Commission, 2018b) mapped out the initial areas in which the Member States should pool their resources and coordinate their actions, the "Coordinated Plan: 2021 Review" (European Commission, 2021b) moves towards an action-oriented approach with concrete joint actions focusing on the implementation of concrete measures and the removal of remaining fragmentation.

It is built around four key pillars: (1) to set enabling conditions for AI development and uptake, (2) to make the EU a place where excellence thrives from the lab to the market, (3) to ensure that AI works for people and society as a force for good, and (4) to build strategic leadership in high-impact sectors. These high-impact sectors encompass areas such as smart mobility, law enforcement, migration and asylum, climate, and the environment. In tandem with the third pillar, which encompasses a promotion of "trustworthy AI" globally and nurturing of talent and skills, it could be seen as a reflection of the second component in the "Communication on AI in Europe," that is the EU's initial AI strategy, which focused on socio-economic changes associated with AI. Although the focus on scaling up the EU technical infrastructure is evidenced across all four pillars, the first is of particular relevance.

The policy document focuses on governance coordination frameworks and, crucially, on data infrastructures and computing capacities in order to create an enabling environment for AI development and uptake.

Referring back to the "European Strategy for Data" (European Commission, 2020a), which aims to establish a single market for data within the EU and the "Proposal for a Regulation on European Data Governance (Data Governance Act)" (European Commission, 2020e), which proposes several regulatory measures to increase society's trust in data sharing, the "Coordinated Plan: 2021 Review" (European Commission, 2021b) outlines a number of core actions for data and an associated cloud infrastructure. These actions include establishing a new European Alliance for Industrial Data, Edge, and Cloud;[20] co-investing with the Member States in common European data spaces; and a European cloud federation and investigating the opportunity to set up an Important Project of Common European Interest (IPCEI) for next generation cloud infrastructures.

Prior to the "Coordinated Plan: 2021 Review" (European Commission, 2021b), at the end of 2020, 27 Member States signed a joint "Declaration on Building the Next Generation Cloud for Businesses and the Public Sector" (European Commission, 2020c) where they expressed their intention to establish a secure, trustworthy, and competitive cloud infrastructure in Europe for public administration, businesses, and citizens alike. It addresses efforts aligned with those in the "Coordinated Plan: 2021 Review" (European Commission, 2021b), such as pooling of EU, national, and private investment; shaping the process in accordance with the European Alliance on Industrial Data and Cloud;[21] and fostering technical solutions and policy norms for an interoperable pan-European cloud service. Adjacent to this is Gaia-X,[22] an independent European platform for cloud infrastructure launched by France and Germany,[23] which intends to increase European cloud competitiveness vis-à-vis the U.S. and China.

With an eye to infrastructure, the "Coordinated Plan: 2021 Review" (European Commission, 2021b) looks to support the development of High-Performance Computing capabilities, as well as AI hardware. The latter encompasses investment in micro-electronics for AI chips, neuromorphic computing, photonics, and projects under the Electronic Components and Systems for European leadership Joint Undertaking (ECSEL JU).[24] More specifically, actions call for the launch of an Industrial Alliance on Microelectronics, supporting research and innovation actions for low-power edge AI, and investing in processor and semiconductor technologies.

In fact, as part of its goal of achieving Digital Sovereignty, the EU quite evidently is aiming to advance its capabilities and to lessen its reliance on international actors when it comes

to the design and production capabilities of low-power processors for AI and towards 2nm processor technologies. In late 2020, 18 Member States signed a "Declaration on a European Initiative on Processors and Semiconductor Technologies" (European Commission, 2020b) to consolidate resources and boost the EU's electronics and embedded systems value chain.

The "Coordinated Plan: 2021 Review" (European Commission, 2021b) also accounts for High-Performance Computing. It encourages Member States to continue developing large-scale High-Performance Computing infrastructure and references the importance of the EuroHPC Joint Undertaking.[25]

A proposed new independent regulation for the EuroHPC[26] is expected to lead to an increase in the acquisition and development of supercomputers in the EU, rebalancing the scale in favor of the EU. Overall, it aims to develop exascale supercomputers with over "1 billion billion" operations per second (10^18 ops/second), to support the development of quantum and hybrid computers (also described in the 2018 regulation, section 21) and to create 33 "national competence centers" which will help to provide easier access to HPC opportunities locally and strengthen knowledge and expertise.

The second pillar in the "Coordinated Plan: 2021 Review" (European Commission, 2021b) focuses on knowledge transfer and horizontal actions to support research and innovation (R&I). Its actions cover stakeholder collaboration, expanding and mobilizing research capacities, building up suitable TEFs, and funding AI solutions and ideas. Stakeholder collaboration will range from the Public–Private Partnership on AI, Data, and Robotics[27] to a co-programmed European Partnership on Photonics,[28] supporting the EU's drive towards technological sovereignty. Whereas the European Commission already invested over €50 million in AI excellence centers through Horizon 2020,[29] it suggests funding more AI excellence centers and encourages Member States individually to set up regional and national excellence centers. Reminding ourselves of this chapter's earlier discussion on the importance of "trustworthy AI" for EU AI governance, it needs to be highlighted that the document suggests that funded programs for AI under Horizon Europe are expected to adhere to the "ethics by design" principle, including "trustworthy AI." We can see that various earlier threads are starting to come together. Finally, as mentioned in the "White Paper on AI" (European Commission, 2020d) and the earlier "Coordinated Plan" (European Commission, 2018b), Digital Innovation Hubs (DIHs) play a role in strengthening the ecosystem. To that end, alongside TEFs for specific sectors, such as edge AI or agri-food, the European Commission will support a scaling up of existing DIHs and set up new networks for what it terms European Digital Innovation Hubs (EDIHs) with AI expertise. These EDIHs will connect SMEs and start-ups with resources made available via the AI-on-Demand platform and relevant TEFs to make the AI system ready for deployment within the EU market. Finally, an investment of €1 billion from Horizon Europe and the Digital Europe programs is expected between 2021–2027. The Digital Europe program overall budget funds artificial intelligence (€2.1bn), HPC (€2.2bn) and cybersecurity (€1.7bn). The ambition remains the same as in the earlier "Coordinated Plan" (European Commission, 2018b), namely to raise this to €20bn per year through public and private investment. Other institutions that are expected to fund AI in the EU are the European Innovation Council,[30] the European Investment Bank[31] (via the European Innovation Fund[32]), and the European Institute of Innovation and Technology.[33]

One issue the EU has historically faced is that promising companies are often purchased by foreign companies before they reach their full potential, gobbling up talent and knowledge

in the process. Although this is not an explicit part of strengthening the EU's AI landscape nor mentioned in the "Coordinated Plan: 2021 Review" (European Commission, 2021b) it is relevant to quickly mention the EU's regulatory framework to screen foreign direct investment (FDI).[34] This FDI framework aims to protect the EU's strategic interests and came into force at the end of 2020. Of particular relevance in light of the EU's shift towards achieving digital sovereignty[35] and scaling its technical infrastructure, is that this framework covers assets that are "critical technologies and dual use items."[36] This includes amongst others AI, robotics and semiconductors.

From the previous discussion is evident that the EU is both (1) serious in its pursuit to strengthen its vision of human-centric "trustworthy AI" by shaping the AI governance framework through regulation, policy and certification; and, (2) willing to build out the entire ecosystem in support of this vision, positioning itself as a future sovereign digital actor and third way between the U.S. and China.

The Future

Sketching the future of AI governance in the EU

We have seen how the EU built up its strategy and how all the elements fit in with the larger tapestry of the EU's approach to AI governance. I will now explore the current dynamic that the EU is exhibiting and briefly sketch three AI governance areas that are prime candidates to become crucial for AI governance in the EU in the coming decade.

The finalization and implementation of the "AI Act" (European Commission, 2021b) over the coming months and years is a clear candidate, but there are other less obvious but equally relevant ones. The following paragraphs pick them out, polish them, and highlight their importance in the future pathways for EU AI governance.[37]

AI megaprojects: A CERN for AI and AI lighthouses

Multiple experts have called for megaprojects within the EU over the past few years. Most notably, for a CERN for AI.[38] Certainly, such a project would be very ambitious. Nevertheless, upon careful reading of recent EU policy documents and the general drive to boost the technical landscape and ecosystem (as previously outlined) a budding AI megaproject may be on the cards.

There are roughly two shapes such a project could take: being centralized within a Member State who has suitable technical infrastructure and is well located or broadly distributed across Member States, with one centralized headquarter. Given the existing ecosystem, ongoing efforts to boost research capacity and technical infrastructure and recent advocacy from large research groups within the EU,[39] both options may be viable.

As described earlier, the EU plans to build out their existing DIHs into EDIHs with significant focus on AI EDIHs. This would lead to a stark increase in the number of facilities where research can be conducted and where AI systems can be developed and experimented upon by SMEs and start-ups. Moreover, the European Commission has recently funded a

large-scale project called ELISE,[40] the European Network of AI Excellence Centres. ELISE collaborates with the European Laboratory for Learning and Intelligent Systems (ELLIS), a large network of European researchers, and closes the gaps between AI institutes in Europe. This adds to a previously funded network, TAILOR[41] (Foundations of Trustworthy AI—Integrating Learning, Optimization and Reasoning), whose goal is to create a network across Europe on the "Foundations of Trustworthy AI." Efforts such as these evidence that there is fertile ground to establish a centralized large-scale headquarter from an increasingly powerful network and quasi-independent nodes.

Another key reason for why a large-scale AI project might be both on the cards and meaningful for the EU to establish can be found in the "AI Act" (European Commission, 2021b). In tandem with new regulatory requirements, more TEFs will be needed. This ranges from TEFs specialized to test and assess for specific aspects of the conformity assessment, as well as those that can assess an AI system's entire regulatory fitness. Building out existing facilities for testing and experimentation and establishing novel ones will eventually lead to a dense landscape of distinct but similar institutions across the EU's Member States. It might be in the EU's best interest to centralize these facilities and locate them alongside big industrial efforts such as the European Cloud efforts, European Data Spaces, and the HPC Joint Undertaking. Such a localization could increase efficiency and provide economies of scale for using data, research engineering, and other supporting infrastructure; enable more ambitious research, testing, and experimentation efforts; and, encourage a laser sharp alignment between policy and practice.

Indeed, the European Commission has indicated ambitions to develop something akin to a CERN for AI in several policy documents, for example in the "White Paper on AI" (European Commission, 2020d), and most recently in the "Coordinated Plan on AI: 2021 Review" (European Commission, 2021b). In particular, the development of AI Lighthouse Centres (or, a center) within the EU is championed. This would be a large-scale research facility for AI.

It would be promising for the EU's ambitions on a global playing field to establish an AI lighthouse center, a CERN for AI, or another version of a large-scale facility for AI research and development. Most importantly, this could lead to the EU becoming a truly unified player where fragmentation between various European research institutes is superseded (Stix, 2018), significant chunks of the aimed for EUR 20bn per year funding for AI could be centralized for ambitious projects, and new talent could be attracted to the EU (Stix, 2019).

Considering a future landscape with an increasing number of networked institutions, mounting calls from large AI research networks within the EU, and the policy proposals from the European Commission, the future of EU AI governance may well hold an AI megaproject.

AI agencies: Regulation, measurement, and foresight

The idea of a large European AI Agency is not new. The very first document presented in this chapter, the resolution on "Civil Law Rules on Robotics" (European Parliament, 2017), already called for the establishment of an EU Agency for Robotics and AI in "order to provide the technical, ethical and regulatory expertise needed to support the relevant public actors, at both Union and Member State level, in their efforts to ensure a timely, ethical and

well-informed response to the new opportunities and challenges, in particular those of a cross-border nature."

Similarly, the 2019 European Parliament report "A Comprehensive European Industrial Policy on Artificial Intelligence and Robotics" (European Parliament, 2019) called for the establishment of a European regulatory agency for AI and algorithmic decision-making. With this backdrop, and tracing the institutional landscape mapped out by the "AI Act" (European Commission, 2021c), it is likely that the EU will eventually establish a new institution, specifically for the governance of AI. The "AI Act" (European Commission, 2021c) envisions a complex institutional interplay to sustain the regulatory measures for AI. This encompasses various national institutions, including those that fall under the National Competent Authorities, which would be the National Supervisory Authority, the Notifying Authority, and various Notified Bodies (official conformity assessment bodies); Market Surveillance Authorities; and, from the European Commission's side, a novel European AI Board (where, for example, Member States will be represented), and an expert group. All of these will play a crucial role for the application of the horizontal regulation for AI within the EU and will have different scopes and powers. Some will have investigative power, and some will assess the suitability of AI systems for the European market. Of course, many of these institutions cannot (and should not) be merged. Nevertheless, after an initial phase of getting to know the ropes of the final agreed upon regulation, it is likely that there will be a time window in which an EU AI Agency would be built to combat fragmentation, pool expertise, and streamline various workflows.

Beyond the aforementioned scopes, the "AI Act" (European Commission, 2021c) also has a provision which ensures that new AI systems can be added to the list of high-risk AI systems as and when deemed appropriate. To ensure that timeliness and foresight are underlining this power, and more generally, to ensure that policy making matches technological progress, I propose that another version of an EU AI Agency—which has not been part of any EU-level discussions yet—should be considered: a European AI observatory.

Historically, observatories were established to measure and survey natural occurrences (e.g., astronomical, geophysical, or meteorological events). An AI observatory as envisaged here, on the other hand, would monitor, measure, and benchmark AI progress, a technology created by humans.

Although the EU is involved in OECD efforts towards an international AI policy observatory and has its own body, the AI Watch,[42] this does not yet live up to what an EU AI observatory could look like. As envisaged here, the EU AI observatory should have the capacity and ambition of conducting independent forecasting and measurement exercises. These in turn would ensure that policy making, regulation, and other governance efforts in the EU are aligned with the technical state of the art of AI systems and sufficiently future proof.

As previously indicated, in order for policy makers and regulators to make suitable and timely decisions to add potential future high-risk AI systems to the "AI Act" (European Commission, 2021c), either to regulate them or to ban them, they need to be aware of ongoing technical developments. One way of doing so from a government's perspective could be to monitor the technical landscape, measure technical progress, and therewith notice crucial shifts that could indicate a cause for concern or intervention.

Overall, AI governance in the EU, through regulation, standards, certification, or other efforts could be significantly more impactful, agile, and anticipatory if it narrowed the pacing gap between technological progress and governance efforts (Marchant et al., 2011).

Furthermore, metrics can be seen as comparatively non-threatening and could encourage information sharing between countries and institutions, indirectly promoting collaboration and cooperation.

Taking these aspects into consideration, I suggest that an EU AI observatory would be vastly beneficial to the EU's AI ambitions and could be seen as a contributing factor for better regulatory measures in the future.

Standards

Lastly, standards will play an important role in the future of the EU's AI governance. While standardization efforts are not a "European-only" effort, they are likely to meaningfully shape the EU market for AI systems. Crucially, the "AI Act" (European Commission, 2021c) notes that if suitable standards exist[43] that would cover one or more of the relevant legal requirements for an AI system to pass conformity assessment, then an adherence to those standards can be considered as an adherence to the legal requirement(s) in question. Of course, if a provider chooses not to follow an existing standard or, where a standard does not exist then they must prove suitable and sufficient adherence to the legal obligations in a different manner. This goes to say that actors involved in standardization bodies and those directly working on standards will have some non-negligible leeway in shaping the future mechanisms with which many of those adhering to the regulatory framework for AI will tackle their conformity assessments. Standardization efforts might frame some of the hurdles a high-risk AI system needs to pass before it enters the EU market and shape the manner in which relevant actors think about assessing high-risk AI systems.[44] I propose that it is a key lever for upcoming governance measures within the EU.

Conclusion

In conclusion, this chapter first introduced the background of the EU's AI governance ambitions, drawing together the different elements and highlighting how they interconnect, together developing the tapestry out of which the EU's vision and current governance efforts result. Subsequently, it introduced the two-pronged recipe the EU pursues for AI. There, it discussed the corresponding roles of "trustworthy AI" and regulatory efforts with the associated scaling of the technical infrastructure within the EU. Together, it demonstrated how these choices support the EU's ambition to be a leader on ethical and human-centric AI on an international stage. Finally, I sketched three possible future directions they envision the EU moving towards: AI research megaprojects, new AI Agencies, and an increasing importance of standardization efforts.

Notes

1. We will revisit the role this Report continues to play later in this chapter.
2. See https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf.

3. Of course, such coordination may be seen as implicit by virtue of these countries being Member States of the European Union.
4. This Declaration has been expanded on significantly in the Coordinated Plan on AI and its follow up.
5. It should be noted that the Declaration is non-binding. However, any other international efforts are equally non-binding at the time of this writing.
6. These will be further explored later in this chapter. Another deliverable not discussed in this chapter are the Policy Recommendations for Trustworthy AI.
7. This includes investment on Member State-level, as well as public–private partnerships, the Digital Europe Programme, and Horizon Europe funding (the latter two both run between 2021–2027).
8. A feedback mechanism was also used in the case of the Assessment List, for which feedback was received through two online questionnaires and through in-depth interviews with different types of organizations.
9. The European Commission opened a broad stakeholder consultation process where feedback was solicited through three different streams: (a) a quantitative stream, (b) a qualitative stream, and (c) a holistic stream. The quantitative stream consisted of two surveys, one for developers and deployers, and one for other stakeholders. The qualitative stream allowed for 50 in-depth day long interviews with selected companies trialing the Assessment List on use-cases. Finally, the last channel allowed for feedback from the broader community, discussion papers, white papers, blog posts, and reports were provided alike from entities as broad as individual researchers to international industry.
10. See https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf.
11. It should be noted that the key requirements and Ethics Guidelines are of a non-binding format.
12. See https://digital-strategy.ec.europa.eu/en/funding/international-alliance-human-centric-approach-artificial-intelligence.
13. See https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf.
14. See https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=PI_COM:Ares(2020)3896535&from=EN.
15. See https://www.europarl.europa.eu/news/en/press-room/20200925IPR87932/making-artificial-intelligence-ethical-safe-and-innovative. The second legislative initiative is on "liability for AI causing damage," focusing on civil liability claims against AI-systems. The third report addresses intellectual property rights (IPRs) with relation to AI, suggesting that AI lacks a legal personality, and therefore inventorship should be exclusive to humans.
16. The proposal for a Regulation on a European approach for Artificial Intelligence will need to pass through European Parliament and the Council. Once these two institutions agree on a final text, the regulation will be adopted.
17. See https://cordis.europa.eu/project/id/825619.
18. Horizon Europe is the current Multiannual Financial Frameworks program and runs from 2021–2027 to support research, science, and innovation with EUR 95.5 bn.
19. See https://digital-strategy.ec.europa.eu/en/activities/digital-programme.
20. See https://digital-strategy.ec.europa.eu/en/library/cloud-and-edge-computing-different-way-using-it-brochure.
21. See https://digital-strategy.ec.europa.eu/en/news/towards-next-generation-cloud-europe.
22. See https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html.

23. See https://www.euractiv.com/section/digital/news/digital-brief-the-gaia-x-generation/?utm_content=1591278775&utm_medium=eaDigitalEU&utm_source=twitte.
24. See https://www.ecsel.eu/what-we-do-and-how#:~:text=The%20ECSEL%20Joint%20Undertaking%20%2D%20the,era%20of%20the%20digital%20economy.
25. See https://eurohpc-ju.europa.eu/.
26. See https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1592.
27. See https://ai-data-robotics-partnership.eu/.
28. See https://www.photonics21.org/#:~:text=The%20European%20Technology%20Platform%20Photonics21,growth%20and%20jobs%20in%20Europe.
29. See https://digital-strategy.ec.europa.eu/en/news/towards-vibrant-european-network-ai-excellence.
30. See https://eic.ec.europa.eu/index_en.
31. See https://www.eib.org/en/index.htm.
32. See https://ec.europa.eu/clima/policies/innovation-fund_en.
33. See https://eit.europa.eu/.
34. See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019R0452.
35. See https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en.
36. As defined in Article 2.1 of Regulation (EC) No 428/2009.
37. It should be noted that this section is the personal opinion of the author and the likelihood of the proposed sketches coming into fruition varies.
38. See https://www.steven-hill.com/why-we-need-a-cern-for-ai/.
39. See https://claire-ai.org/wp-content/uploads/2020/02/CLAIRE-Press-Release-11.pdf; https://www.timeshighereducation.com/news/scientists-split-europe-paves-way-cern-of-ai; https://sciencebusiness.net/news/call-cern-ai-parliament-hears-warnings-risk-killing-sector-over-regulation.
40. See https://cordis.europa.eu/project/id/951847.
41. See https://liu.se/en/research/tailor/about.
42. A joint initiative between the European Commission's Joint Research Centre (JRC) and the Directorate General for Communications Networks, Content and Technology (DG CONNECT). See more: https://knowledge4policy.ec.europa.eu/ai-watch/about_en.
43. Those standards would have to be published in the Official Journal of the European Union.
44. Assuming the provider chooses to use standards or technical specifications for their conformity assessment. However, if standards are available it is likely that most providers will choose to adhere to the standards to streamline and minimize their workflows between various geographical regions.

References

European Commission. (2018a). *Communication from the Commission—Artificial Intelligence for Europe* (COM(2018) 237 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

European Commission. (2018b). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—Coordinated Plan on Artificial Intelligence* (COM/2018/795 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN.

European Commission. (2018c). *(Digital Day) Declaration on Cooperation on Artificial Intelligence*, European Commission website—JRC Science Hub—Communities. https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/eu-declaration-cooperation-artificial-intelligence.

European Commission. (2019). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions—Building Trust in Human-Centric Artificial Intelligence* (COM/2019/168 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168.

European Commission. (2020a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions—A European strategy for data* (COM(2020) 66 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0066&from=EN.

European Commission. (2020b). *Declaration A European Initiative on Processors and semiconductor technologies*. European Commission website—Library. https://digital-strategy.ec.europa.eu/en/library/joint-declaration-processors-and-semiconductor-technologies.

European Commission. (2020c). *Declaration—Building the next generation cloud for businesses and the public sector in the EU*. European Commission website—News & Views. https://digital-strategy.ec.europa.eu/en/news/towards-next-generation-cloud-europe.

European Commission. (2020d). *On Artificial Intelligence—A European approach to excellence and trust* (COM(2020) 65 final). European Commission. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Commission. (2020e). *Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act)* (COM/2020/767 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767.

European Commission. (2020f). *Report from the Commission to the European Parliament, the Council, and the European Economic and Social Committee—Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics* (COM/2020/64 final). European Commission. https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX:52020DC0064.

European Commission. (2021a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions—Fostering a European approach to Artificial Intelligence* (COM/2021/205 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:205:FIN.

European Commission. (2021b). *Coordinated Plan on Artificial Intelligence 2021 Review*. European Commission. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

European Commission. (2021c). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (COM/2021/206 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

European Economic and Social Committee. (2017). *Opinion of the European Economic and Social Committee on 'Artificial intelligence—The consequences of artificial intelligence on the*

(digital) single market, production, consumption, employment and society' (own-initiative opinion) (2017/C 288/01. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016IE5369.

European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and 'autonomous' systems*. Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382.

European Parliament. (2017). *European Parliament resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics* (2018) (2015/2103(INL)) (2018/C 252/25). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051&qid=1620812299497.

European Parliament. (2019). *Report on a comprehensive European industrial policy on artificial intelligence and robotics* (2018/2088(INI)). European Parliament website. https://www.europarl.europa.eu/doceo/document/A-8-2019-0019_EN.html.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. https://doi.org/10.2139/ssrn.3518482.

Hagendorff, T. (2019). The ethics of AI ethics—An evaluation of guidelines. *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1903.03425.

Independent High-Level Expert Group on Artificial Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for trustworthy AI. European Commission. AI HLEG. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. Springer Science & Business Media.

Schiff, D., Biddle, J., Borenstein, J., *&* Laas, K. *(2020). What's next for AI ethics, policy, and governance? A global overview*. *In 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375627.3375804*.

Stix, C. (2018). The European AI landscape. *Workshop Report*. European Commission. DG Connect. http://ec.Europa.Eu/newsroom/dae/document.Cfm.

Stix, C. (*2019). An infrastructural framework to achieve a European artificial intelligence megaproject*.

Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics 27* (1), 15.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking artificial intelligence principles. *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1812.04814.

CHAPTER 47

AI AND INTERNATIONAL POLITICS

AMELIA C. ARSENAULT AND SARAH E. KREPS

Introduction

In recent years, the fields of artificial intelligence (AI) and machine learning have seen significant advancements, including improvements in pattern recognition, predictive analytics, and machine vision, that have greatly accelerated global interest in these technologies. Bolstered by the increased volume and accessibility of data, AI has been heralded for its ability to efficiently synthesize information and improve decision-making processes. Given these contemporary advancements and their broad applications for a range of sectors, a number of states have begun investing heavily in AI, including China, the United States, Israel, the United Kingdom, Canada, and Russia, amongst others. While the United States has maintained its lead in AI technology, and has claimed that it intends to protect its competitive edge, other states also recognize the important role that AI is likely to play in the future. For example, Russian President Vladimir Putin famously stated that whoever leads in AI will "become ruler of the world," and China has also expressed a desire to become a world leader in AI by 2030.

In light of contemporary technological advancements in the fields of AI and machine learning, coupled with the global investment in research and development, scholars have begun considering the impact that this technology will have on international and domestic politics. While it seems increasingly certain that AI will feature prominently in global affairs, academics disagree about the precise ways in which AI will influence international politics. Some argue that AI will produce radical and rapid transformations (Dafoe, 2018), while others argue that the effects of AI will be gradual and incremental (Levy, 2018), thereby allowing the international community to avoid some of its potentially destabilizing effects. Others predict that AI will fundamentally shift the balance of power in the international sphere, providing smaller states and non-state actors with unprecedented opportunities to increase their power and expand their scope of influence. Alternatively, many scholars maintain that Chinese and American strength in AI technology will not fundamentally alter the global balance of power but will instead entrench bilateral competition (Mialihe,

2018). In relation to military applications of AI, scholars are divided over whether or not the pursuit of this technology will prompt a global arms race, undermine stability, and increase the likelihood of military escalation (Johnson, 2019; Johnson, 2020).

This chapter starts from the premise that AI is best understood as an accelerating and enabling force. As a result, the role that AI is playing and will play in international relations suggests that it is less likely to produce drastic, unforeseen transformations in domestic and international politics as it is to accelerate and exacerbate trends that were already underway. Many of the potential consequences of AI, including disinformation and growing polarization, weapons automation, income inequality, the erosion of liberal democratic norms, and surveillance, are not unique to the effects of AI or the direct consequences of AI proliferation. Instead, the rush to acquire and harness the power of AI has further established and intensified these trends—allowing disinformation to be increasingly targeted and precise, aiding the development of weapons that operate with higher degrees of autonomy, worsening income inequality, and further expanding opportunities for surveillance. However, the fact that AI is an enabling force does not suggest that it is trivial or inconsequential. Indeed, while the particular consequences that AI holds for global politics remain to be seen, and many of the aforementioned debates are speculative, it is clear that AI is unlikely to fade from the realms of global governance and international affairs. As AI continues to improve, and states capitalize on the opportunities for efficient governance, military leverage, and global power and influence, it is likely to have significant implications for both domestic and global politics as an accelerator and enabler. As such, it is critical that international actors, academics, and the general public more broadly begin preparing for the era of AI.

This chapter begins by outlining a number of potential applications of AI for international politics, from military and defense, to trade and diplomacy. It then examines the reasons why international actors are likely to continue pursuing AI in ways that influence politics and concludes by evaluating leading debates about the future of global politics in a world where many actors harness this emerging technology.

AI: Current Applications for International Politics

AI is being integrated into a range of sectors that have important implications for international politics. First, many states have begun investing in military applications of AI, including the U.S., China, Russia, the United Kingdom, France, Israel, and South Korea (PAX, 2019). The integration of AI into military and defense, particularly in terms of the development of lethal autonomous weapons, has garnered significant backlash from a number of international actors, including state representatives and non-governmental organizations (NGOs). Critics have expressed concern about the moral, ethical, and legal implications of lethal, AI-powered weapons in relation to maintaining respect for human dignity, enforcing accountability for casualties, discriminating between non-combatants and legitimate targets, and the need to retain the distinctively "human" nature of warfare (Sparrow, 2007; Sharkey, 2010). Importantly, platforms need not employ AI to operate with

a degree of autonomy, as weapons developers have long sought to automate both offensive and defensive capabilities. For example, weapons such as the Israeli Harpy and Harop loitering munitions, which use sensors to independently identify and target enemy radar, operate with high levels of autonomy because they do not require direct human intervention to select targets (Rogoway, 2016). Further, the South Korean SGR-A1 robot uses lasers, heat sensors, voice recognition, and machine vision to identify individuals crossing the demilitarized zone between North and South Korea (Lin et al., 2008, p. 19). The SGR-A1 robot is also equipped with machine guns, and is capable of autonomous targeting (Lin et al., 2008, p. 19; Crootof, 2015, p. 1869). Both of these weapons retain "human-in-the-loop" capabilities, whereby a human operator can intervene to override the system in case of inappropriate use; in the South Korean case, representatives for Samsung have refuted reports certifying the SGR-A1's high level of autonomy, maintaining that it can only be used in a way that requires human intervention (Caron, 2020, p. 174).

Nonetheless, the significant investment into military AI, coupled with the potential for a "first-mover" advantage has raised concerns about the progressive development of increasingly lethal autonomous weapons that operate with little human oversight. For example, the Turkish military announced in June 2020 that it would be receiving 500 KARGU kamikaze drones fitted with artificial intelligence and facial recognition capabilities. These drones, previously deployed to monitor the Turkish border with Syria, have prompted concern from military observers and NGOs about the increasingly prominent role of advanced AI in weapons systems (Hambling, 2020; PAX, 2019b). A recent UN report notes that soldiers were "hunted down and remotely engaged" by KARGU-2 drones in Libya (United Nations Security Council, 2021, p. 17). The report also notes that the weapons were "programmed to attack targets without requiring data connectivity between the operator and the munition," prompting debate about the level of autonomy granted to the weapons and raising questions about whether there were opportunities for human oversight (United Nations Security Council, 2021, p. 17; Cramer, 2021). While it is unclear whether anyone died as a result of the strikes, or how much human intervention ensued, the use of KARGU-2 drones in Libya is likely a harbinger of things to come (Hernandez, 2021).

AI also has important military and defense applications beyond the development of autonomous weapons systems. For example, the U.S. Behavioural Learning for Adaptive Electronic Warfare (BLADE) program uses machine learning to analyze transmissions and communications and is pursuing initiatives to "quickly detect and characterize threat transmissions, automatically synthesized waveforms optimized to jam the detected signals, and analyze its own effectiveness against these signals in the field" (Kenyon, 2011). The Pentagon's Project Maven relies on machine learning of drone footage to classify and identify items of interest, including potential targets (Horowitz et al., 2018). AI's predictive capabilities can also be integrated into military planning, as demonstrated by the U.S. Marines' use of AI to measure military readiness; using data on individual units and their capabilities, this tool can predict which battalions should be deployed, and can analyze how different configurations of resource allocation could improve readiness (Corrigan, 2018). In terms of predictive analytics, U.S. forces have used Palantir's predictive algorithms to assist in predicting the location of IEDs in Afghanistan and Iraq. This technology has also been integrated into security efforts and counter-terrorism initiatives in Denmark and Israel (Winston, 2018; Peretti, 2017).

AI also plays an increasingly prominent role in strategic communications between international actors. Importantly, disinformation is not a new tactic, but has long been used by adversaries to smear opponents, generate confusion, and weaponize information (Posetti & Matthews, 2018). However, through the use of psychometrics, botnets, sentiment analysis, and generative neural networks, AI has allowed disinformation efforts to be more precise, curated, and destabilizing. Bolstered by the immense amount of data available from social media platforms, AI-facilitated disinformation is capable of generating predictions about the types of content that is most likely to resonate with a particular audience (Horowitz et al., 2018). For example, Cambridge Analytica used machine learning to predict social media users' personality types and preferences in the lead up to the 2016 Brexit referendum and 2016 American Presidential election; through the use of AI-facilitated psychometrics, these algorithms then allowed for the micro-targeted, "highly personalized" dissemination of disinformation based on individual preferences and biases (Manheim & Kaplan, 2019; Brkan, 2019; Kertysova, 2018). AI-powered bots are also capable of rapidly and autonomously flooding communications with a particular narrative or false story in order to influence public opinion, drown out competing narratives, and sow doubt (Kreps, 2020; Kendall-Taylor et al., 2020).

Most recently, advances in machine learning and natural language processing have led to the rise of "deepfake" technology, which uses neural networks and machine learning to create image, text, and audio forgeries that seem genuine (Schippers, 2020). By exploiting the natural human inclination that "seeing is believing," these forgeries may have significant consequences for politics, particularly when used to depict politicians in unfavourable ways. For example, a doctored video purporting to show Nancy Pelosi, Democratic Speaker of the United States House of Representatives, drunkenly slurring during a speech was circulated throughout social media in 2019 (Schippers, 2020, p. 33). While the video was later debunked as a fake, it was retweeted and spread by prominent politicians and Pelosi's political adversaries, including former President Donald Trump and his lawyer, Rudy Giuliani (Forgey, 2019). Political parties have also used "deepfake" technology to depict foreign leaders. In May 2018, the Belgian Flemish Socialist Party shared a fake video that featured President Trump chastising Belgium for remaining in the Paris climate agreement (Von Der Burchard, 2018). While the video in question was easily debunked as fake, the willingness of a political party to use this technology to gain political leverage demonstrates its implications for international relations. This technology is becoming increasingly accessible, prompting concerns about further opportunities for blackmail, deniability, and the erosion of truth (Kertysova, 2018).

AI also offers unique opportunities for states to censor dissent, spread propaganda, and monitor opposition. Governments have long engaged in repression and surveillance before the advent of AI, with states using radio and the Internet to manipulate public opinion and monitor communications (Dragu & Lupu, 2021). However, AI has greatly intensified surveillance capabilities, rendering them more all-encompassing and thorough. For example, China has integrated AI into their surveillance apparatus, including "high-resolution cameras, facial recognition, spying malware, automated text analysis, and big-data processing" (Kendall-Taylor et al., 2020). Most notably, China has relied heavily on digital repression in the Xinjiang region, using facial recognition technology to monitor and surveil members of the Uyghur community. China has collected and analyzed swaths of individual data regarding "cell phone information, genetic data, and information about

religious practices" in order to generate predictions, monitor behavior and identify who shall be sent to "re-education" camps (Kendall-Taylor et al., 2020). In recent years, the export of Chinese surveillance technology to a range of other countries, including Malaysia, Myanmar, Ethiopia, Venezuela, and Zimbabwe, has driven concerns about the rise of "digital authoritarianism" globally, whereby the sale of advanced technologies may contribute to the repression of civil liberties, the monitoring of activists, and the strengthening of authoritarian regimes (Andersen, 2020; Feldstein, 2019). While analyses of the ways AI might facilitate repression and expansive surveillance have largely focused on authoritarian states, it is important to note that the associated threats to individual rights also apply to the liberal democratic context. Democracies may be less likely to use AI to engage in the type of egregious repression most commonly associated with authoritarian regimes; however, the integration of these technologies into specific sectors, particularly policing, the judicial system, and border security, risks undermining the liberal democratic norms of privacy, non-bias, and transparency and further entrenching systemic injustices.

While the aforementioned applications of AI establish its connection to insecurity and repression, AI may also facilitate international cooperation and negotiation. AI is already being used for simple tasks relating to diplomacy and consular work, including the processing of passport and visa applications, refugee aid, and consular registrations (Bjola, 2020). As natural language processing and predictive analytics continue to improve, AI could facilitate diplomacy by reducing language barriers, contributing to the administration of humanitarian aid to developing states, and allowing for the monitoring of political trends to improve development and peacekeeping (Horowitz et al., 2018). AI is also being incorporated into international trade, as exemplified by the International Chamber of Commerce (ICC)'s use of the Cognitive Trade Advisor (CTA), an AI system which is capable of rapidly analyzing existing trade agreements in order to provide negotiators with relevant information to improve future trade negotiations (Bjola, 2020).

Motivations and Incentives for International Actors Pursuing AI

Many states are dedicating resources to research, development, and investment into emerging AI technologies. The contemporary boom in AI investment and research by a wide range of international actors necessitates an analysis of exactly why state and non-state actors alike are motivated to harness this technology. We outline three primary incentives that international actors have for pursuing AI: improving domestic governance and efficiency, capitalizing on emerging military opportunities, and pursuing global power and influence.

Improving domestic governance and efficiency

Firstly, international actors face strong incentives to pursue AI given the associated opportunities it offers for increased efficiency, improved organization, and problem-solving.

With its ability to synthesize and analyze information with precision, AI can recognize patterns and trends likely to go unnoticed by human analysts, and organize vast amounts of information (Ünver, 2018). As such, this technology is being used to identify patterns and generate predictions for a range of sectors including healthcare, finance, agriculture, law, education, and policing, amongst others (König & Wenzelburger, 2020, p. 6). In fact, the rise of AI and predictive algorithms has led to a growing embrace of "data-based" or "evidence-based" policies by a range of states, whereby existing data is used to generate predictions about future trends, preferences, and patterns. For example, many states have embraced predictive policing initiatives (Yen & Hung, 2021; McCarthy, 2019) which purport to analyze data about the characteristics and areas affiliated with crime in order to develop pre-emptive, risk-based responses. Despite evidence that predictive policing programs exacerbate racial bias in the criminal justice system and may threaten the liberal democratic norms and values of accountability and transparency, municipal governments have lauded advanced algorithms as allowing for more efficient, "data-based" crime policy.

Predictive capabilities are also being used to identify vulnerabilities in emergency preparedness and disaster response, thereby allowing international actors to make necessary changes to critical infrastructure, or otherwise prepare for emergencies (Lu et al., 2021). "Smart city" initiatives, embraced by cities as diverse as Singapore, Barcelona, New Delhi, Tel Aviv, New York, and Montreal use sensors, facial recognition, and other information and communication technologies to collect data on local activity. When fed into predictive models, this data allows a range of city services, from waste removal, air quality, lighting, and transportation, to policing and emergency response, to adapt to citizen needs in real-time in an attempt to improve the provision of local services, resource allocation, and sustainability (Estrada et al., 2019, p. 12; Tironi & Criado, 2015, p. 90). Thus, international actors have real incentives to pursue emerging AI technology in an attempt to improve domestic governance, reduce costs, and facilitate enhanced organization.

International actors also have a financial incentive to pursue AI. AI is often lauded for its ability to increase efficiency by reducing the burden placed on human analysts (Horowitz et al., 2018). That is, processes that use machine learning can be retrained with "reinforcement learning," thereby allowing the system to detect its own mistakes and hypothetically allowing for increased accuracy and efficiency (Levy, 2018). In sectors where employment involves significant danger or repetition, replacing certain tasks with automation and smart manufacturing may also allow firms to reduce operating costs associated with employment and worker error, as well as human costs related to injury (Ünver, 2018). By eliminating the potential for emotion, exhaustion, and other distinctly human qualities that may interfere with the completion of routine tasks, AI might also increase productivity and reduce worker error (Ünver, 2018). It is important to note that the economic consequences of AI will not be realized for some time, and as such, are hard to fully predict.

However, should the commercial sector continue to lead the development of AI technologies, the resulting innovation could also benefit state economies more broadly by bolstering productivity and investment (Furman & Seamans, 2018). While the adoption of automation will likely exacerbate income inequality, particularly for those completing highly repetitive tasks, and may harm those firms incapable of incorporating these emerging technologies, it has been estimated that AI could add $15.7 trillion to the world economy in 2030, with $6.6 trillion emanating from gains in productivity, and $9.1 trillion from consumption (PricewaterhouseCoopers, 2017). Consequently, individual firms as well as states

that aim to increase their share of global GDP are likely to encourage innovation and development in emerging technological sectors, including AI and machine learning.

Capitalizing on military opportunities

Secondly, international actors are pursuing AI in an attempt to leverage the potential military opportunities associated with increased automation, improved prediction, and anomaly-detection. Neural networks and machine learning allow for improved situational awareness, particularly when used to analyze drone feeds and images to identify patterns and detect anomalies. Military programs that embrace AI-facilitated image analysis, such as Project Maven, can therefore be used to "optimize battle plans" (Horowitz et al., 2018, p. 48). For those international actors capable of harnessing this emerging technology, the integration of AI into weapons systems may also offer distinct military opportunities.

Those that support the development of AI-powered weapons, capable of identifying and engaging a potential target independently, often argue that these weapons will increase the speed and precision of targeting and reduce the risk to human operators. For example, small, autonomous, AI-powered drones, capable of communicating and coordinating with each other, may allow militaries to operate in areas that were previously inaccessible, allowing for stealth, speed, and swarming (Garcia, 2019; Johnson, 2020; Horowitz et al., 2018). The integration of AI and robotics automation into weapons systems allows militaries to circumvent many of the human frailties that limit the efficacy of military operations, including emotion, fatigue, and the hesitation to engage targets. AI will also play a critical role in cyberwarfare and defense, as AI can be used to identify system vulnerabilities and rapidly recognize suspicious cyber activity, malware, or hacks. It is important to note, however, that just as this technology will allow for rapid detection of cyber anomalies and threats, it will also improve the offensive cyber capabilities of nefarious actors that aim to disrupt, monitor, or degrade an adversary's cyber network (Johnson, 2020).

AI can also be integrated into military and defense in ways that are not directly linked to warfare or conflict. For example, AI that integrates big-data and predictive modeling can make inferences about scheduling, equipment and weapons maintenance, psychological distress, and military medicine and healthcare, thereby improving military readiness (Johnson, 2020). Palantir, a leading data analytics firm that has previously worked with a number of state militaries and defense agencies, promises to assist "military decision makers leverage critical information for data-driven decisions" and "rapidly turn mountains of data into plans of action" (Palantir Defense, n.d.). Just as non-military sectors and industries are increasingly replacing tedious labour with automation, AI may be capable of completing repetitive tasks, thereby allowing personnel to take up more productive work (Etzioni & Etzioni, 2017; Jensen et al., 2020, p. 539). AI can also support military intelligence by identifying trends, which can then be used to predict behaviour and highlight anomalous or suspicious information (Horowitz et al., 2018).

Albeit for different reasons, both democracies and autocracies have interests in pursuing military applications of AI. Faced with the possibility of electoral backlash following battlefield casualties, democracies may be especially incentivized to embrace military applications of AI that reduce the risk to soldiers. This reduced risk may also be achieved by the integration of predictive analytics that purport to improve military decision-making

by identifying crucial vulnerabilities. By minimizing the human costs associated with engaging in conflict, democratically-elected leaders can shield themselves from the electoral consequences associated with military casualties. Similarly, autocratic leaders that face credible threats from their own military may be able to use AI to "keep tabs on government officials, gauging the extent to which they advance regime objectives and rooting out underperforming officials" (Kendall-Taylor et al., 2020). In times of increased tension between a leader and military elites, access to AI may therefore allow authoritarian leaders to further consolidate personal power (Horowitz, 2018). Thus, it is likely that both democratic and authoritarian states will pursue AI in order to access the associated military opportunities.

Power, influence, and global competition

The ability of international actors to acquire and harness emerging technologies has long been linked to perceptions of identity, influence, and global power. AI is therefore simply the latest innovation with important implications for global competition, where "powerful technology has often been a driving force in the pursuit of a world order... the nation-state with the technology in question is in a privileged position to pursue its objectives" (Ramamoorthy & Yampolskiy, 2018, p. 3). Given its wide applicability to a range of industries and sectors, including military and defense, competency in the AI arena is increasingly understood as representative of state power.

As a lucrative, critical tool for 21st century governance, the pursuit of cutting-edge developments in AI is a high priority issue for many states that strive to play a significant role in international affairs. The view that these capabilities are emblematic of global power has contributed to a fear of "falling behind," as exemplified by rhetoric expressing the critical importance of remaining abreast of the most recent technological developments (PAX, 2019). Arguing that the U.S. needs to develop a competitive, comprehensive AI strategy, former Deputy Defense Secretary Bob Work noted that "if we wait for the Sputnik moment, we will be too far behind" (PAX, 2019, p. 34). In light of evidence that EU states had been lagging behind in AI patents and startups, European leaders have since expressed ardent commitment to bolstering their technological capabilities and emphasizing the importance of crafting ethical guidelines for AI (Brattberg et al., 2020, p. 5; Candelon & Di Carlo, 2020, Miailhe, 2018).

Importantly, actors will vary considerably in their abilities to harness this technology, with some emerging as developers and innovators, and others largely relying on exports. However, it is clear that many states have a real interest in taking advantage of the opportunities associated with AI, where the successful integration of this technology may serve as a signal to other international actors and competitors that a state is capable of adapting and embracing emerging technological trends. As a consequence of the rising importance placed on the ability to innovate and develop in this space, the proliferation of AI has encouraged global competition between international actors. Given its association with military readiness, effective domestic governance, and economic strength, AI development is increasingly seen as a "zero-sum game" (You & Dingding, 2018, p. 257). The ways in which the quest for AI dominance has spurred competition between actors is best elucidated by rising tensions between the United States and China. Nevertheless,

the significant consequences that AI holds for economic growth, military capabilities and perceptions of power indicate that this technology is likely to continue influencing global competition between a range of state actors.

Just as international actors are enticed by the potential "hard power" opportunities associated with military AI, the pursuit of emerging technologies can also be used to bolster "soft power" by strengthening an actor's "indirect cultural, commercial, and political influence" globally (Mialhe, 2018, p. 12). Importantly, the concept of "soft power," in contrast to more traditional characterizations of power that emphasize "hard" military and economic capabilities, refers to an actor's ability to influence others, and to incentivize other actors to behave in ways that are favorable without resorting to coercion or the threat of force (Nye, 2004, p. 256). Actors are also likely to use research and education in order to expand their "soft power"; while both China, the U.S., and others continue to pursue research and development in this field in order to bolster their soft power and political influence, the U.S. is currently the world leader in AI research (PAX, 2019; You & Dingding, 2018, p. 246). AI-powered disinformation efforts, whereby actors influence their adversary's political environment not through coercion or the threat of force but through the subtle manipulation of information, can also be used as a tool of "soft power" (Kamarck, 2018).

In terms of economic and cultural influence, China has exported AI technology globally, sustaining their reputation as technological leaders in the process (Polyakova & Meserole, 2019). Chinese tech sales have often been framed as an effort on behalf of the CCP to "export authoritarianism," whereby other authoritarian states emulate China's own use of advanced technology to repress and surveil (Polyakova & Meserole, 2019; Wright, 2018). As previously mentioned, evidence suggests that authoritarian states, including Ecuador, Zimbabwe, the United Arab Emirates, Venezuela, and Pakistan have been some of the most prominent importers of Chinese-made surveillance technology. For example, the 2019 purchase of Chinese facial recognition systems and other smart city technology by Serbia has triggered condemnation from local civil liberties advocates who argue that the system has already been used to target protestors and has had a chilling effect on freedom of speech (Higgins, 2021; Stojkovski, 2019). Further, the deep linkages between the Chinese state and domestic tech companies has prompted concern that these exports could allow the CCP to access sensitive data about foreign citizens (Arthur, 2018). The sale of advanced technologies may therefore allow AI innovators such as China to subtly influence the political preferences of other actors, increasing economic interdependence and expanding their sphere of influence.

However, those countries that do purchase AI-based technologies from China are not necessarily destined to use it for repression and undue surveillance. The ways that these technologies are used will vary and will depend on a range of domestic factors and local circumstances (Feldstein, 2020; House Committee Hearing on "China's Digital Authoritarianism: Surveillance, Influence, and Political Control," 2019, p. 3). Further, a number of liberal democracies, including the Netherlands, Germany, Japan, Italy, and France have purchased smart city technology from China, an anomalous trend that seems to counter the "digital authoritarianism" narrative (Australian Strategic Policy Institution & International Cyber Policy Centre, 2019). Whether those states that purchase AI-powered technologies from China will use them to further consolidate authoritarian power and repress dissent remains to be seen, but China's innovation in AI promises to expand its global influence as these countries become reliant on this technology (and its inevitable upgrades).

AI and Implications for International Politics

Given the continued development and proliferation of AI technology, scholars have begun considering the potential consequences that AI may have for global politics. Below we consider three central ways that AI may influence contemporary international politics: altering global trends in authoritarianism and liberal democracy, shifting the global balance of power, and influencing modern warfare and conflict.

Global trends in authoritarianism and liberal democracy

Primarily, AI may affect global trends in authoritarianism and liberal democracy. As previously mentioned, AI is likely to exacerbate the scope and scale of disinformation as psychometrics, predictive algorithms, and "deepfake" technology become accessible and more sophisticated in their ability to mimic human interaction (Kertysova, 2018; Horowitz et al., 2018; Schippers, 2020). Importantly, liberal democracies are particularly vulnerable to disinformation, and also face distinct challenges in addressing or mitigating its effects. Democratic participation requires that citizens are capable of accessing credible information in order to encourage political debate and the formulation of political opinion. As "deepfake" technology, targeted disinformation, and botnets become capable of sowing confusion and informational chaos, these efforts promote distrust of traditional media. Should citizens gain awareness of the prominence of AI-facilitated disinformation, including "deepfakes," there is a risk that they will become overly skeptical of all news that does not confirm to their own beliefs, allowing for further disengagement with traditional sources and the reinforcement of "echo chambers." Rising distrust of traditional media sources, coupled with the dramatic rise in the number of alternative "news" sources made available through online platforms may insulate individuals from differing viewpoints, thereby exacerbating the intensity of pre-existing biases and generating polarization. Further, as botnets become increasingly sophisticated in their ability to simulate human communications, their prominence in political debates may make certain ideologies or opinions seem more popular than they really are, thereby distorting political debate and influencing perceptions of general sentiment (Kreps, 2020, p. 20). Democratic discourse is therefore at risk of being eroded as individuals are unable to share a basic standard of "truth" regarding political issues and events.

Further, liberal democracies emphasize the importance of freedom of speech and expression. As such, they are fundamentally limited in their ability to curb disinformation, even when it emanates from foreign sources and involves malicious propaganda or patently false stories, because efforts to address disinformation may be characterized as an infringement of freedom of speech (Tenove, 2020, p. 531). In cases where the disinformation being disseminated is highly critical of liberal values, efforts to "debunk" falsities may, ironically, strengthen those very narratives by prompting "individuals to double down on incorrect information" (Jankowicz, 2020, p. 202). Disinformation, in its ability to exploit existing divisions, solidify echo chambers, and intensify polarization, threatens to erode

the discourse, institutional trust, and access to credible news essential to the functioning of democracy.

Importantly, the undermining of liberal democracy does not only result from AI's use by nefarious actors, whether foreign or domestic. Instead, democracies must also ensure that the integration of this technology into their own domestic governance reflects liberal democratic values. Many liberal democratic states have embraced predictive policing, which relies on algorithms to evaluate the likelihood that a particular individual, group, or area will be associated with criminality. However, research has demonstrated that many predictive algorithms reflect a distinct racial bias, where algorithms that are trained on data that reflects the disproportionate over-policing and incarceration of Black and Latinx individuals will reify and reproduce these inequalities in their predictions (Bacchini & Lorusso, 2018; Garvie & Frankle, 2016). Biased results may therefore be framed as "objective," "data-based" calculations, but they always reflect the social, political, and economic context from which the data is derived. Further, facial recognition technology has been found to be unable to properly differentiate between Black individuals, further compounding racial bias (Schippers, 2020).

Given the increased interest in integrating predictive AI into law enforcement, the judicial system, healthcare, and social services, liberal democracies are at risk of further entrenching and reifying existing societal injustices in ways that are fundamentally antithetical to core democratic values. Further, algorithmic decisions often remain opaque, whereby the processes through which the algorithm produces predictions are difficult to understand, including for those with sophisticated tech backgrounds (Young et al., 2021). This issue, commonly referred to as the "black box" problem, undermines the norms of democratic accountability and transparency, as citizens who do not sufficiently understand the algorithms that govern their daily lives cannot be reasonably expected to hold their leaders accountable for their misuse (Schippers, 2020; König & Wenzelburger, 2020). Lacking explainability, decisions made using AI become inscrutable (Young et al., 2021, p. 12). Indeed, the embrace of AI may lead to quantification bias, where numerical data is "overvalorized" so that decisions made using advanced algorithms are "uncritically assigned credence on the basis of this perceived expertise" (Young et al., 2021, p. 10). The proliferation of AI may also undermine the liberal democratic values of freedom of assembly and speech; as facial recognition and predictive analytics become an increasingly popular tool for monitoring protests and online communications, citizens of democracies may become more hesitant to demonstrate or express their more controversial opinions online (Swan, 2020).

In addition to the challenges that AI poses for liberal democracy, AI may also offer opportunities for authoritarian regimes to further consolidate power. By adopting predictive algorithms and AI-embedded surveillance systems, leaders may be able to reduce their reliance on military elites, thereby increasing their personalist power. Further, AI-powered surveillance allows for more invasive, thorough, and rapid monitoring of citizen behaviour compared to the capabilities of human agents, facilitating a more expansive and efficient surveillance regime.[1] More broadly, the increased use of advanced communications technologies also allows authoritarian regimes to monitor, harass, and coerce diaspora communities or foreign dissidents living abroad through the hacking of personal devices (Deibert, 2020). Authoritarian regimes may also use facial recognition technologies to identify protestors or dissidents participating in public demonstrations; knowledge that

one's government has access to this technology may have a "chilling" effect on the incentive to protest (Knaus, 2018; Murphy, 2018, p. 26). In fact, Kendall-Taylor et al. (2020) find that autocrats who use forms of digital repression, including facial recognition and surveillance, face lower rates of protest than those who do not. Lastly, autocrats are also capable of using bots, psychometrics, and microtargeting to sow doubt, disseminate propaganda, and "shape public perception of the regime and its legitimacy" (Kendall-Taylor et al., 2020; Horowitz et al., 2018).

Some of the potential consequences of AI's global proliferation have implications that would affect both authoritarian and democratic states in similar ways. Namely, should further development prompt significant transformations in social, political, and economic relations at the domestic level, particularly in terms of rising unemployment and lowered wages, AI could generate instability for both authoritarian and democratic states (Dafoe, 2018; Horowitz et al., 2018; Zeng, 2020). As many lose their jobs to automation, economic and social changes initiated by the expansion of AI into many sectors may increase inequality, thereby increasing support for populist or nationalist movements that can use automation as a framework to levy attacks against the "elites" (Levy, 2018). Changes to the labour market can also influence politics and social relations, whereby those that are disadvantaged by economic changes grow disenfranchised, resentful, and opposed to their domestic leadership. Rising hostility towards automation and the "elites" could produce instability in both democratic and authoritarian regimes, with Horowitz et al. (2018, p. 19) noting that previous disruptions to the labour force have historically triggered "mass turmoil, coups, and other tensions." In terms of general trends in regime types and the erosion of liberal democratic norms, AI could increase political conflict in ways that make "domestic unrest, insurgencies, civil wars, nationalism, xenophobia, and a turn to authoritarianism more likely" (Horowitz et al., 2018, p. 15).

However, domestic instability would perhaps be more damaging for individual authoritarian leaders and parties unaccustomed and unprepared for flux and challenges. For example, Zeng (2020, p. 1455) notes that the destabilizing social and political effects of AI may be particularly consequential in China, where mass unemployment or rising income inequality amongst the large civilian population could lead to unrest that risks the CCP's stability and legitimacy. That is, if AI contributes to rising inequality and opposition, individual authoritarian leaders are at risk of being ousted; however, at a broader international level, major social, political, and economic transformations enabled by AI could contribute to the erosion of liberal democracy, whilst bolstering authoritarianism.

AI does, however, offer opportunities for improving and strengthening democracy. If designed in such a way that checks against systemic injustices and instead prioritizes equality and access, AI could be used to improve access to services and resources. For example, AI could improve education by allowing for more personalized, adaptive, and inclusive lesson plans (Vincent-Lancrin & van der Vlies, 2020; Luckin et al., 2016). Further, AI's ability to rapidly analyze and synthesize large amounts of data could improve access to information, thereby empowering individuals with timely and relevant knowledge about the political environment (Schippers, 2020). Access to reliable information may then allow citizens to "keep track of political actors' performance" and hold their elected leaders accountable (König & Wenzelburger, 2020, p. 5). AI could also be used to strengthen communication between elected leaders and their constituents, allowing politicians to better understand voters' interests and demands (König & Wenzelburger, 2020). Just as AI could

be used to exacerbate disinformation and information echo chambers, it could also connect like-minded citizens, allowing them to collectively mobilize in the pursuit of a shared political cause (Schippers, 2020).

AI could also be used to defend against efforts to undermine democracy. For instance, researchers have begun using AI to detect deepfakes and AI-fabricated text in an attempt to reduce the spread of disinformation. There are significant limitations to this approach, as the technology used to detect disinformation often lags behind the newest tactics employed by adversaries. However, AI could still be useful for detecting and classifying disinformation attempts in some instances. As previously mentioned, AI is also being used to support cyber security, namely through the use of anomaly-detection and machine learning to rapidly identify and respond to adversarial intrusions. As cyber-attacks are becoming increasingly sophisticated and have begun targeting critical infrastructure, malicious cyber actors could target democratic election systems (Hodgson et al., 2020, p. 3; Cybersecurity and Infrastructure Security Agency, 2020). Should AI prove useful in identifying "zero-day" cyber threats targeting election processes, it could defend against attacks on democracy.

AI can therefore improve access to information and critical services, facilitate communication, and strengthen defenses to bolster democracy. However, the realization of these potential benefits depends upon the precise ways in which AI is developed and used. The current state of AI development in liberal democratic states reflects fundamentally asymmetric power relations, where the power to create and use these systems has been consolidated by a small group of individuals with interests that are distinct from those who stand to be most affected by AI systems. In order for AI to strengthen, rather than undermine, democracy, it must be developed in ways that recognizes potential risks, reduces systemic biases, and ensures that AI systems serve the people who are subjected to them. Elected leaders and private firms alike should not frame AI as a "silver bullet" technology that can replace human subjectivity with innate objectivity. Instead, a more democratic approach to AI adoption requires recognition that some decisions should never be made by algorithms and that some tasks require uniquely human characteristics such as empathy and an understanding of historical, social, and political context. The pursuit of AI does not necessarily entail the demise of democracy. However, if AI is to support and strengthen democracy, it must emphasize respect for a common humanity and be framed as a supportive tool to be leveraged in certain circumstances, as opposed to an objective response to a host of societal and political problems.

Balance of power

Whether AI will influence the global balance of power is a source of scholarly debate. Focusing on the military applications of AI, Horowitz (2018, p. 54) argues that it is innately difficult to predict how AI will influence global power relations because its effects on organizational structures and institutions remain unseen (Horowitz, 2020). Conversely, some scholars maintain that AI is likely to transform which actors control key economic assets, surveillance and repressive capabilities, and information, thereby influencing the global balance of power (Dafoe, 2018). As previously mentioned, China and the U.S. are currently the world leaders in emerging AI technologies, leading to predictions that AI competition

will exacerbate "an overall trend towards the centralization of power in the hands of a few actors" (Mialihe, 2018).

While the U.S. currently retains a "first-mover advantage" and supremacy in AI, including military applications, China is quickly closing the gap, as exemplified by growing rhetoric regarding the importance of tech dominance, the rise of domestic tech companies, and considerable investment initiatives (Johnson, 2019, p. 2; Kanaan, 2020). Importantly, the Chinese structure of central economic planning allows it to access private data harvested by domestic firms with few regulatory limitations, thereby posing a direct threat to the technological dominance of the US (Zeng, 2020). However, the U.S. benefits from a commercial industry that is conducive to innovation, access to a large reserve of talent, and an abundance of data on "logistical planning, promotion and assignments, and operations on the front lines" that may give them an advantage in military prediction (Horowitz, 2020; You & Dingding, 2018). While it is difficult to predict precisely how AI will alter global power balances, it is likely that current AI "superpowers" will continue to compete for dominance in emerging tech markets.

The proliferation, and increased accessibility of AI technology may also influence the global balance of power by destabilizing bipolar competition between the U.S. and China, instead allowing a wider range of actors, including non-state actors, to enhance their relative power (Horowitz, 2018). It is important to note that states with smaller budgets and limited investment in research and development will face barriers to entry, given the high costs associated with the necessary software, hardware, and computing power needed to harness advanced algorithms and predictive capabilities (Horowitz, 2018; Horowitz, 2020). However, should the costs associated with this technology continue to fall, and AI continue to be integrated into a range of sectors and industries, it will become more accessible for states with smaller budgets, as well as non-state actors. Horowitz (2020) notes that research has demonstrated that technologies that can be used in military settings but can more broadly be described as "general-purpose," referring to their wide applicability to a range of sectors, tend to proliferate faster than technology that is solely used for military purposes. As a general-purpose technology, information regarding AI development is not shrouded in military secrecy and concerns about national security, thereby creating opportunities for a wider range of actors to copy, mimic, or steal emerging innovations. The diffusion of AI technology to smaller state actors is also facilitated by trade from large actors. As previously mentioned, China has exported AI technology widely, including facial recognition and surveillance technology, to both authoritarian and democratic states alike.

While non-state actors face significant barriers in harnessing AI technology, they are expected to attempt to exploit the organizational and military advantages it offers. Scholars have noted that just as terrorist groups were capable of accessing drones, which they have used for surveillance, reconnaissance, and combat, they may be able to acquire autonomous weapons through illegal purchases or leaks from commercial industries (Allen & Husain, 2018; Ware, 2019). Gartenstein-Ross et al. (2019) argue that if a particular AI system is "commercially available, cheap, and easy to use," as well as suitable for an individual group's "strategies, planning structures, and recruitment tactics," non-state actors are likely to attempt to adopt the system in question (p. 60). While they note that adoption does not inherently spell successful usage, they predict that violent non-state actors are likely to pursue "deepfake" technology, virtual agents or "chatbots," social network mapping and sentiment analysis, as well as object recognition capabilities in an attempt to strengthen recruitment

efforts, identify targets, engage in cyber intrusions, and stoke tensions (Gartenstein-Ross et al., 2019).

Non-state actors may also be able to use AI to exacerbate the damaging effects of cyber-attacks, where machine learning reduces the need for actors to have a sophisticated understanding of cyber, thereby "lowering the bar for attacks by individuals and non-state groups and increasing the scale of potential attacks for all actors" (Horowitz et al., 2018, p. 4). Importantly, state actors have begun expressing concerns regarding the implications that non-state access to AI could hold for the global balance of power. For example, the U.S. 2018 National Defense Strategy notes that the spread of AI technology "means that state competitors and nonstate actors will also have access... a fact that risks eroding the conventional overmatch to which our nation has grown accustomed" (Department of Defense, 2018). In addition to bolstering their own power relative to other actors, non-state access to AI will also lack regulation and oversight (Ramamoorthy & Yampolskiy, 2018).

Rising global interest in AI technologies has also bolstered commercial firms, which have been the primary entities at the forefront of cutting-edge technological progress in AI and machine learning. Large tech companies, responsible for creating powerful algorithms and holding huge swaths of data about individual preferences and behaviour, have therefore played an increasingly important role in global politics in recent years. While the number of firms with significant global influence is relatively small, these entities are capable of using their access to data in order to "further entrench their market position and monopolies," supporting the continued "centralization of digital power" (Johnson, 2019a; Mialhe, 2018, p. 5). Further, AI innovations made in the commercial sector are being applied to the military and defense sectors, facilitating the growing influence of said firms and representing a reversal of historical trends whereby military technologies were refitted for civilian use (Missiroli, 2020). The immense power of these companies has prompted debates over the past several years regarding their relative lack of regulation. As the entities primarily responsible for breakthroughs in AI technology, these firms are primarily concerned with profit and the acquisition of data, generating risks for accountability, transparency, and oversight (Ünver, 2018). That is, there are concerns that the lucrative profit opportunities associated with AI innovation will create an incentive for these firms to innovate without full consideration of AI's ethical and normative consequences (Ramamoorthy & Yampolskiy, 2018).

Warfare and conflict in the age of AI

In recognizing the many military and defense-related applications for AI, scholars have begun considering the implications that AI will have on conflict and warfare. Firstly, there is concern that as AI is increasingly perceived as a critical tool for contemporary governance as well as an indicator of status and power, actors will eagerly pursue this technology without sufficiently considering the need for comprehensive regulation and safety standards. That is, the ambition to exploit the opportunities associated with this technology may trigger a global arms race, where access to this technology is prioritized over safety and the creation of reasonable limitations on its use. As this technology continues to proliferate, it may also alter the decision to use force. Should military AI allow for an offensive advantage, either through autonomous weapons or the use of predictive analytics and image recognition, this

technology could prompt actors to strike preemptively, thereby reducing the threshold for conflict and increasing the likelihood of military escalation (Johnson, 2019b).

AI-powered military capabilities that allow for an offensive advantage could therefore raise the risk of conflict (Dafoe, 2018). Johnson (2020) argues that the global proliferation of AI could increase the risk of nuclear war, where the increased speed of AI-facilitated warfare would severely undermine stability. Further, the introduction of increased automation and AI into military affairs raises important questions about attribution and responsibility. The use of AI in warfare, particularly in terms of cyberwarfare and autonomous weapons where it is difficult to attribute responsibility, increases deniability, further undermining the ability for international actors to effectively respond to attacks (Missiroli, 2020). Consequently, the ability to claim plausible deniability and the challenges associated with concretely attributing AI-facilitated strikes could increase the likelihood of force and pre-emptive strikes (Scharre, 2019) as actors can rest assured that they are not likely to face retribution (Missiroli, 2020).

Despite their potential advantages, AI-enabled weapons also raise moral and legal challenges associated with the integration of AI into warfare. Primarily, weapons systems that incorporate AI into surveillance or targeting aim to use data and advanced algorithms in order to improve the precision of strikes. However, if systems that use AI are not capable of sufficiently adapting to "changing circumstances," perhaps due to a lack of relevant data, the ability to integrate information quickly in order to produce a response could result in escalation, "false alarms, mistakes, or accidents" (Garcia, 2019, p. 2). Military AI is also currently incapable of appreciating the distinctly human social and political contexts in which it operates, which may also lead to civilian deaths and unlawful targeting. Importantly, autonomous weapons have an unclear application to international law (Castel & Castel, 2016). While international humanitarian law requires that the responsibility for conflict and violence not be "transferred to inanimate beings," initiatives to integrate military AI into international law have been undermined by the global community's inability to agree over exactly what entails an "autonomous" weapon (Garcia, 2019, p. 3; Missiroli, 2020). International law has therefore not yet been successful in efforts to preemptively regulate AI-powered weapons before they become advanced and widely accessible.

While an abundance of literature on the ways AI will influence warfare and conflict has largely focused on the risks of a global arms race, escalation of threat, and instability, AI may also offer opportunities for improving global cooperation and coordination. As previously mentioned, AI is already being integrated into trade negotiations in order to rapidly provide negotiators with an abundance of information required to facilitate additional trade deals. Further, AI's impressive translation capabilities may also improve diplomacy and cooperation by reducing language barriers between actors, thereby facilitating communication (Horowitz et al., 2018). In a similar fashion to the use of AI to summarize important information in existing trade negotiations, AI can also monitor treaties for "accuracy, consistency, and completeness," thereby potentially helping actors overcome commitment problems (Bjola, 2020, p. 12; Dafoe, 2018, p. 20). Machine learning could also be used to predict trends and potential outcomes including those related to "social or political tensions" to improve preventative measures and allow for the rapid deployment of appropriate aid (Bjola, 2020, p. 18).

Conclusion

As AI continues to advance and proliferate, it is likely to influence relations between international actors in a range of ways, from military planning, trade negotiations, and strategic communications. While scholars diverge in their predictions of the ways in which this technology will influence international politics, it is important to recall that AI as a technology is not innately authoritarian and repressive, nor is it inherently conducive to democratic governance. Instead, the basic technology and algorithms behind AI simultaneously create opportunities for improved international coordination and cooperation between actors, and exacerbate risks of invasive surveillance, competition, and escalation of military tensions. As opposed to producing a radical reconfiguration of international and domestic politics, it seems that the proliferation of AI has accelerated and solidified existing trends. It is therefore critical that scholars attempting to understand the ways in which AI is currently influencing international relations and predict its future implications for global order emphasize the distinct political, social, and economic contexts in which these implications arise. As an increasingly critical tool of contemporary governance, AI will play a central role in future relations between international actors; it is therefore incumbent upon scholars, states, and the global community more broadly to begin preparing for international politics in the era of AI.

Note

1. It is important to note, however, that some scholars have challenged the assumption that AI will disproportionately benefit authoritarian regimes. For example, Farrell, Newman, and Wallace (2022) argue that authoritarian governments may have to rely on low-quality data as citizens censor their true opinions and preferences, ultimately undermining predictive accuracy.

References

Allen, J., & Husain, A. (2018). AI will change the balance of power. *U.S. Naval Institute Proceedings 144* (8), 26–31.

Andersen, R. (2020). The panopticon is already here. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/.

Arthur, H. (2018). Huawei's (and China's) dangerous high-tech game. *Forbes*. https://www.forbes.com/sites/arthurherman/2018/12/10/huaweis-and-chinas-dangerous-high-tech-game/?sh=67726baa11ab.

Australian Strategic Policy Institution & International Cyber Policy Centre. (2019). Mapping China's tech giants. https://chinatechmap.aspi.org.au/.

Bacchini, F., & Lorusso, L. (2018). Race, again: how facial recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society 17* (3), 321–335.

Bjola, C. (2020). *Diplomacy in the age of artificial intelligence* (Emirates Diplomatic Academy Working Paper). Emirates Diplomatic Academy. https://static1.squarespace.com/static/

52c8df77e4b0d4d2bd039977/t/5e3a9a45d29b7f336bbda061/1580898895236/EDA+Working+Paper_Artificial+Intelligence_EN+copy.pdf.

Brattberg, E., Csernatoni, R., & Rugova, V. (2020). *Europe and AI: Leading, lagging behind, or carving its own way?* (Carnegie Endowment for International Peace Working Paper) Carnegie Endowment for International Peace. https://carnegieendowment.org/files/BrattbergCsernatoniRugova_-_Europe_AI.pdf

Brkan, M. (2019). Artificial intelligence and democracy: The impact of disinformation, social bots and political targeting. *Delphi-Interdisciplinary Review of Emerging Technologies 2* (2), 66–71.

Candelon, F., & di Carlo, R. C. (2020). Europe is missing out on the A.I. revolution—but it isn't too late to catch up. *Fortune*. https://fortune.com/2020/12/04/europe-ai-competitiveness-gap/.

Caron, J. (2020). Defining semi-autonomous, automated and autonomous weapon systems in order to understand their ethical challenges. *Digital War 1*, 173–177.

Castel, J. G., & Castel, M. E. (2016). The road to artificial super-intelligence: Has international law a role to play? *Canadian Journal of Law and Technology 14* (1), 1–15.

Corrigan, J. (2018). Marines turn to artificial intelligence to better deploy troops. *Nextgov*. https://www.nextgov.com/emerging-tech/2018/11/marines-turn-artificial-intelligence-better-deploy-troops/153151/.

Cramer, M. (2021). A.I. drone may have acted on its own in attacking fighters, U.N. says. *The New York Times*. https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html.

Crootof, R. (2015). The killer robots are here: Legal and policy implications. *Cardozo Law Review 36*, 1837–1915.

Cybersecurity and Infrastructure Security Agency. (2020). *Election infrastructure cyber risk assessment*. (Critical Infrastructure Security and Resilience Note). Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/sites/default/files/publications/cisa-election-infrastructure-cyber-risk-assessment_508.pdf.

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

Department of Defense. (2018). *Summary of the 2018 National Defense Strategy of The United States of America: Sharpening the American Military's Competitive Edge*.

Deibert, R. (2020). *Reset: Reclaiming the Internet for civil society*. House of Anansi.

Dragu, T., & Lupu, Y. (2021). Digital authoritarianism and the future of human rights. *International Organization 75* (4), 991–1017.

Estrada, E., Martinez Vargas, M. P., Gomez, J., Perez Negron, A. P., Lopez, G., & Maciel, R. (2019). Smart cities big data algorithms for sensor location. *Applied Sciences 9*, 1–14.

Etzioni, A., & Etzioni, O. (2017). Pros and cons of autonomous weapons systems. *Military Review 97*, 72–81.

Farrell, H., Newman, A., & Wallace, J. (2022). Spirals of Delusion: How AI distorts decision-making and makes dictators more dangerous. *Foreign Affairs* 101 (5), 174–181.

Feldstein, S. (2020). When it comes to digital authoritarianism, China is a challenge—but not the only challenge. *War On The Rocks*. https://warontherocks.com/2020/02/when-it-comes-to-digital-authoritarianism-china-is-a-challenge-but-not-the-only-challenge/.

Feldstein, S. (2019). The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy 30* (1), 40–52.

Forgey, Q. (2019). Guiliani appears to defend sharing a doctored Pelosi video. *Politico*. https://www.politico.com/story/2019/05/24/rudy-giuliani-pelosi-video-1343262.

Furman, J., & Seamans, R. (2018). *AI and the economy* (NBER Working Paper 24689). National Bureau of Economic Research Working Papers. https://www.nber.org/system/files/working_papers/w24689/w24689.pdf.

Garcia, E. V. (2019). *AI & global governance: When autonomous weapons meet diplomacy*. United Nations University Centre for Policy Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3522616.

Garvie, C., & Frankle, J. (2016). Facial-recognition software might have a racial bias problem. *The Atlantic*. https://apexart.org/images/breiner/articles/FacialRecognitionSoftwareMight.pdf.

Gartenstein-Ross, D., Shear, M., & Jones, D. (2019). *Virtual plotters. Drones. Weaponized AI?: Violent non-state actors as deadly early adopters. Valens Global International Strategies & Security*, 1–73.

Hambling, D. (2020). Turkish military to receive 500 swarming kamikaze drones. *Forbes*. https://www.forbes.com/sites/davidhambling/2020/06/17/turkish-military-to-receive 500-swarming-kamikaze-drones/?sh=4fbcoceb251a.

Hernandez, J. (2021). A military drone with a mind of its own was used in combat, U.N. says. *NPR*. https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d.

Higgins, A. (2021). Serbia hails Chinese companies as saviors, but locals chafe at costs. *The New York Times*. https://www.nytimes.com/2021/03/27/world/europe/serbia-chinese-companies-pollution.html.

Hodgson, Q. E., Brauner, M. K., & Chan, E. W. (2020). *Securing U.S. elections against cyber threats: Considerations for supply chain risk management*. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA512-1.html.

Horowitz, M. C. (2020). *AI and the diffusion of global power*. Centre for International Governance Innovation. https://www.cigionline.org/articles/ai-and-diffusion-global-power.

Horowitz, M. C., Allen, G. C., Saravalle, E., Cho, A., Frederick, K., & Scharre, P. (2018). *Artificial intelligence and international security*. Center for a New American Security. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-AI-and-International-Security-July-2018_Final.pdf?mtime=20180709122303&focal=none.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review 1* (3), 37–57.

U.S. House of Representatives Permanent Select Committee on Intelligence Hearing on "China's Digital Authoritarianism: Surveillance, Influence, and Political Control," 116th Cong. (2019). https://docs.house.gov/meetings/IG/IG00/20190516/109462/HHRG-116-IG00-Wstate-ChenWeissJ-20190516.pdf.

Jankowicz, N. (2020). *How to lose the information war: Russia, fake news, and the future of conflict*. Bloomsbury Publishing Plc.

Jensen, B., Whyte, C., & Cuomo, S. (2020). Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence. *International Studies Review 22* (3), 526–550.

Johnson, J. S. (2020). Artificial intelligence: A threat to strategic stability. *Strategic Studies Quarterly 14* (1), 16–39.

Johnson, J. (2019). The end of military-techno *Pax Americana?* Washington's strategic responses to Chinese AI-enabled military technology. *The Pacific Review*, 1–28.

Johnson, K. (2019a). AI ethics is all about power. *Venture Beat*. https://venturebeat.com/2019/11/11/ai-ethics-is-all-about-power/.

Johnson, J. (2019b). Artificial intelligence & future warfare: Implications for international security. *Defense & Security Analysis 35* (2), 147–169.

Kamarck, E. (2018). Malevolent soft power, AI, and the threat to democracy. *Brookings*. https://www.brookings.edu/research/malevolent-soft-power-ai-and-the-threat-to-demo cracy/.

Kanaan, M. (2020). *T-minus AI: Humanity's countdown to artificial intelligence and the new pursuit of global power*. BenBella Books.

Kendall-Taylor, A., Frantz, E., & Wright, J. (2020). The digital dictators: How technology strengthens autocracy. *Foreign Affairs*. https://www.foreignaffairs.com/articles/china/2020-02-06/digital-dictators.

Kenyon, H. (2011*)*. DARPA's "smart jammer" to throw the enemy off balance. *Defense Systems*. https://defensesystems.com/articles/2011/01/04/darpa-develops-smart-jammer.aspx.

Kertysova, K. (2018). Artificial intelligence and disinformation. *Security and Human Rights 29* (1), 55–81.

Knaus, C. (2018). Facial image matching systems risks "chilling effect" on freedoms, rights groups say. *The Guardian*. https://www.theguardian.com/world/2018/nov/07/facial-image-matching-system-risks-chilling-effect-on-freedoms-rights-groups-say.

König, P., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly 37* (3), 1–11.

Kreps, S. (2020). *Social media and international relations*. Cambridge University Press.

Levy, F. S. (2018). Computers and populism: Artificial intelligence, jobs and politics in the near term. *Oxford Review of Economic Policy 34* (3), 393–417.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. US Department of Navy, Office of Naval Research. https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac.

Lu, S., Christie, G.A., Nguyen, T.T., Freeman, J.D., & Hsu, E.B. (2021). Applications of artificial intelligence and machine learning in disasters and public health emergencies. *Disaster medicine and public health preparedness* (pp. 1–8). doi:10.1017/dmp.2021.125

Luckin, R., Wayne, H., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Manheim, K., & Kaplan, L. (2019). Artificial intelligence: Risks to privacy and democracy. *Yale Journal of Law and Technology 21* (106), 106–188.

McCarthy, J. O. (2019). AI & global governance: Turning the tide on crime with predictive policing. United Nations University Centre for Policy Research. https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html#info.

Mialihe, N. (2018). The geopolitics of artificial ntelligence. The return of empires? *Politique Étrangère* 3, 105–117.

Missiroli, A. (2020). Game of drones? How new technologies affect deterrence, defence and security. *NATO Review*. https://www.nato.int/docu/review/articles/2020/05/05/game-of-drones-how-new-tec hnologies-affect-deterrence-defence-and-security/index.html.

Murphy, J. R. (2018). Chilling: The constitutional implications of body-worn cameras and facial recognition technology at public protests. *Washington and LEE Law Review 75* (1), 1–32.

Nye, J.S. (2004). Soft power and American foreign policy. *Political Science Quarterly 119* (2), 255–270.

Palantir Defense. (n. d.). Palantir Defense. Palantir. https://www.palantir.com/search.

PAX. (2019). *State of AI: Artificial intelligence, the military, and increasingly autonomous weapons*. Pax for Peace. https://paxforpeace.nl/media/download/state-of-artificial-intelligence--pax-report.pdf.

PAX. (2019b). *Slippery slope: The arms industry and increasingly autonomous weapons*. Pax for Peace. https://paxforpeace.nl/media/download/pax-report-slippery-slope.pdf.

Peretti, J. (2017). Palantir: The 'special ops' tech giant that wields as much real-world power as Google. *The Guardian*. https://www.theguardian.com/world/2017/jul/30/palantir-peter-thiel-cia-data-crime-police.

Polyakova, A., & Meserole, C. (2019). *Exporting digital authoritarianism*. Brookings Institution. https://www.brookings.edu/wp-content/uploads/2019/08/FP_20190827_digital_authoritarianism_polyakova_meserole.pdf.

Posetti, J., & Matthews, A. (2018). *A short guide to the history of "fake news" and disinformation*. International Centre for Journalists. https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf.

PricewaterhouseCoopers. (2017). Sizing the price: What's the real value of AI for your business and how can you capitalise? *PricewaterhouseCoopers*. https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.

Ramamoorthy, A., & Yampolskiy, R. (2018). Beyond mad?: The race for artificial general intelligence. *ITU Journal 1*, 1–8.

Rogoway, T. (2016). Meet Israel's "suicide squad of self-sacrificing drones". *The Drive*. https://www.thedrive.com/the-war-zone/4760/meet-israels-suicide-squad-of-self-sacrificing-drones.

Schippers, B. (2020). Artificial intelligence and democratic politics. *Political Insight 11* (1), 32–35.

Scharre, Paul. (2019). Killer Apps: The Real Dangers of an AI Arms Race. *Foreign Affairs*, May/June 2019.

Sharkey, N. (2010). Saying "no!" to lethal autonomous targeting. *Journal of Military Ethics 9* (4), 369–383.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy 24* (1), 62–77.

Stojkovski, B. (2019). Big brother comes to Belgrade. *Foreign Policy*. https://foreignpolicy.com/2019/06/18/big-brother-comes-to-belgrade-huawei-china-facial-recognition-vucic/.

Swan, B. W. (2020). Feds comb Facebook to hunt down alleged rioters and looters. *Politico*. https://www.politico.com/news/2020/06/12/facebook-riot-loot-justice-department-314 567.

Tenove, C. (2020). Protecting democracy from disinformation: Normative threats and policy responses. *The International Journal of Press/Politics 25* (3), 517–537.

Tironi, M., & Criado, T. (2015). Of sensors and sensitivities: Towards a cosmopolitics of "smart cities"? *Tecnoscienza 6* (1), 89–108.

United Nations Security Council. (2021). Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council. https://digitallibrary.un.org/record/3905159?ln=en.

Ünver, A. (2018). *Artificial intelligence, authoritarianism and the future of political systems*. *Center for Economics and Foreign Policy Studies*, https://www.jstor.org/stable/pdf/resrep26084.pdf?refreqid=excelsior%3A139bfef9d504893729f961f554a57ab4.

Vincent-Lancrin, S., & van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: promises and challenges (OECD Education Working Paper No. 218). OECD Education Working Paper Series 1–17.

Von Der Burchard, H. (2018). Belgian socialist party circulates "deep fake" Donald Trump video. *Politico*. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-be lgian-socialist-party-circulates-deep-fake-trump-video/.

Ware, J. (2019). Terrorist groups, artificial intelligence, and killer drones. *War on the Rocks*. https://warontherocks.com/2019/09/terrorist-groups-artificial-intelligence-and-killer-drones/.

Winston, A. (2018). Palantir has secretly been using New Orleans to test its predictive policing technology. *The Verge*. https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new orleans-nopd.

Wright, N. (2018). How Artificial Intelligence Will Reshape the Global Order. *Foreign Affairs*. https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order.

Yen, C. P., & Hung, T. W. (2021). Achieving equity with predictive policing algorithms: A social safety net perspective. *Science and Engineering Ethics*, *27*(36), 1–16.

You, W., & Dingding, C. (2018). Rising Sino–U.S. competition in artificial intelligence. *China Quarterly of International Strategic Studies 4* (2), 241–258.

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K. C. (2021). Artificial intelligence and administrative evil. *Perspectives on Public Management and Governance 4* (3), 1–15.

Zeng, J. (2020). Artificial intelligence and China's authoritarian governance. *International Affairs 96* (6), 1441–1459.

CHAPTER 48

THE CRITICAL ROLES OF GLOBAL SOUTH STAKEHOLDERS IN AI GOVERNANCE

MARIE-THERESE PNG

Introduction

This chapter examines critical and cross-geographic perspectives of harm reduction strategies in AI governance. It does so by synthesizing scholarship and policy materials, as well as elements of reflexive observation from the author as a practitioner bridging the tensions of Global North and Global South debates across multiple stakeholder groups. This chapter calls for those working in AI governance, as well as relevant areas of international trade law, intellectual property, technical standards and certification, and human rights, to substantively engage with elements of Global South discourses that are in tension with the dominant discourse.

It aims to present a landscape of AI governance for and from the Global South—advanced by critical and decolonial-informed practitioners and scholars—and contrast this with the dominant AI governance discourse led out of Global North institutions. By doing so, it identifies gaps in the dominant AI governance discourse around interpretations of justice, rights, and geopolitical representation, bridging these gaps with separate but relevant discourses of technology and power, localization, and reparation led by Global South-aligned thinkers and actors. By contrasting these two discourses, this chapter discusses what key differences and tensions might mean substantively for the building of AI governance processes. It argues that structural reform is the substantive outcome of stakeholder heterogeneity, and necessary if the risks of AI systems are to be meaningfully mitigated at the societal and geopolitical levels.

This chapter opens describing the growing popularity of inclusive AI governance, introducing the paradox of participation—wherein inclusion can exist while structural harms persist—and challenging the ahistoric characterization of Global South actors as

peripheral to governance processes. It then presents a brief digest of AI "for and from the Global South," enumerating several critical concerns expressed by Global South discourses but neglected by the dominant AI discourse led by Global North institutions. These critical concerns include digital sovereignty as relevant to low and middle-income countries, infrastructural and regulatory monopolies, harms associated with the labour and material supply chains of AI technologies, beta-testing, and commercial exploitation.

The following section contrasts some key tensions between the Global South and dominant Global North AI discourses. It argues that Global South actors should play a key role in AI governance institutions' ability to effectively mitigate risks within an international context and proceeds to propose three roles of Global South actors: (1) as challenging functions to exclusionary governance mechanisms; (2) providing legitimate expertise in the interpretation and localization of risks, which includes a whole-systems and historic view; and (3) providing a source of alternative governance mechanisms; for example, South–South solidarity, co-governance, democratic accountability, and a political economy of resistance.

The final section of this chapter proposes that key differences between the Global South and dominant Global North discourses can be explained in part by historic power dynamics. To progress towards inclusive or participatory governance, this chapter asserts that global AI governance initiatives would need to understand historic structural power inequalities, and restructure processes in ways that redistribute resource, agenda-setting, and decision-making power. Here, this chapter describes the coloniality of power in AI governance and recasts popular AI Governance frameworks, such as the Fourth Industrial Revolution, in a historic light.

Terminology: Global South and North

Both the Global South and Global North are heterogeneous. Given the complex plurality of Global South stakeholders, it is crucial to examine the limitations and utility of "Global South" and "Global North," as much of current literature on unequal distribution of AI risks tend to use this generalized framework. On one hand, this terminology is a useful unifier for solidarity-building across the Global South; on the other, it omits heterogeneity and internal incongruence of Global South AI development strategies and discourse.

Since the end of the Cold War, the Global North has been associated with stable and affluent states and economies, whereas the term "Global South" has referred to economically disadvantaged nation states. A more nuanced approach to the Global South also understands it as the deterritorialized geography of capitalism's externalities (Mahler, 2017), imprinted with well-studied colonial legacies. Global South perspectives center the displaced priorities, concerns, and voices of the global majority (Campbell-Stephens, 2020). As Singh and Guzmán (2021) articulate "we treat 'Global South' as an imperative to focus on cognate lived experiences of the excluded, silenced, and marginalized populations as they contend with data and AI on an everyday basis."

The dominant AI discourse is spearheaded by the Global North—namely industrialized and wealthy state actors in Western Europe and North America, and, within the growing multi-stakeholder governance model, industry, standards setting organizations, and military research/funding bodies. It is also a deterritorialized reproduction of political,

epistemic, economic, and moral hierarchies promulgated during European colonization. As Glissant and Dash (1999) puts it, "The West is not in the West. It is a project, not a place."

The Global "Souths" (Connell, 2007; Comaroff & Comaroff, 2012) are highly heterogeneous, presenting divergent "political regimes, levels of development, ideologies, and geopolitical interests" (Weiss, 2016) which engender regional contestation, and set real limitations to coordination and collective mobilization. The AI discourse from the Souths "operate on a wide spectrum between optimism of leapfrogging and digital transformation of societies on one end and the pessimism of human suffering caused by new forms of data capitalism and colonialism on the other" (Singh, 2021). In addition, agendas between the Global South and North are not to be seen as inherently dichotomous or antagonistic. The North/South binary does not account for subjugated peoples within the borders of wealthier countries, and vice versa—"economic Souths in the geographic North and Norths in the geographic South" (Mahler, 2017).

There are also many countries who "occupy an interstitial position between North and South"; for example, within the "Global East," "those countries and societies that occupy an interstitial position between North and South" (Muller, 2018). North/South binaries are further challenged by Chinese government and industry leadership in AI governance and applied R&D—with pervasive tech industry actors, such as BATX (Baidu, Alibaba, Tencent, Xiaomi), Huawei, JD, and ByteDance driving market competing with US-based GAFAM firms (Google, Amazon, Facebook, Apple, Microsoft). For example, Chinese investment, research, production, and standards-setting power inevitably shapes dominant discourse, with repercussions to the Global South, with indeterminate long-term outcomes. China's Digital Silk Road, under the Belt and Road Initiative, has been characterized along a spectrum of neocolonialism in the Global South on one end (Kleven, 2019), to a viable South–South alternative to Western domination on the other (Sheng & Nascimento, 2021) given China's history of invasion by Western governments and status in Third World politics (Nash, 2003).

The Growing Popularity of "Inclusive AI Governance"

There is an increasing consensus that global governance is experiencing a "crisis of confidence" surrounding its "legitimacy, transparency, accountability and equitable representation" (Chatham House, 2021).

Despite efforts towards inclusion, and advocacy for representation within international agenda-setting, decision-making, and accountability mechanisms in international fora such as the United Nations, it is clear that representational inclusion is the first step of a long project of governance and institutional reform. The aspiration for inclusive governance to mitigate the replication of unequal and harmful geopolitical dynamics, or secure the social benefits of emerging technologies is not new, with "similarities with developments in other emerging technologies characterized by prominent impacts and uncertainties" (Ulnicane et al., 2021).

The growing popularity of "Inclusive AI Governance" has emerged from the advocacy of Global South aligned or allied actors within high-level AI governance forums—here meaning those from the Global South or North who identify the systemic disadvantages of exclusionary dynamics for the Global South. It has also been prompted by the visibility of work by Jobin et al. (2019), who review the global landscape of AI ethics guidelines, drawing attention to the underrepresentation of Africa, South and Central America, and Central Asia in AI ethics debates, wherein "more economically developed countries are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and the demands of global fairness." Similarly, Jasanoff and Hurlbut (2018) remind us to query "who sits at the table, what questions and concerns are sidelined and what power asymmetries are shaping the terms of debate." Eugenio Vargas Garcia (2021) goes on to explicitly ask -"Where is the Global South?" within the international governance of AI.

High-level initiatives who have adopted inclusive multi-stakeholder approaches to AI governance include the United Nations Secretary General's Office of Digital Cooperation (now Technology Envoy Office); Global Partnership on AI (GPAI) in partnership with the International Development Research Centre (IDRC); Organisation for Economic Co-operation and Development (OECD); the United Nations Educational, Scientific and Cultural Organization (UNESCO); World Economic Forum (WEF); and Institute of Electrical and Electronics Engineers (IEEE). They aim to integrate Global South voices and perspectives in the formation of "global guidelines, standards, and frameworks on responsible and ethical AI." These initiatives fall under the rubrics of "AI for All," "AI for Good," or "AI for SDGs," which though constructive are critiqued for their limited and vague definitions of what "social good" means. As Green (2019) notes - "good isn't good enough." Further, as observed by Ulnicane et al. (2021), the idea of inclusive governance may have normative appeal, but it "is not specific about addressing some well-known challenges of the proposed governance mode such as risks of capture by vested interests or difficulties to achieve consensus."

Inclusive governance initiatives have thus far failed to identify adequate methodologies and protocols to redress power imbalances and substantively engage with underrepresented stakeholder groups. Such methodologies are a prerequisite to designing protective guardrails that center the demands of most impacted communities, and do not stifle their efforts of resistance design and self-determination (Costanza-Chock, 2020). In addition, these initiatives rarely engage in a historical-geopolitical analysis of inclusion/exclusion dynamics, and internal institutional structures which distribute risks unevenly across regions, societies, and communities. For example, the term "interdependence" has grown in use over recent decades, describing multi-stakeholder international cooperation "in the face of an increasingly complex and globalising world order" (Keohane & Nye, 1989; Coate et al., 2015). The slogan "The age of digital interdependence," popularized by the United Nations Secretary General's Office of Digital Cooperation (2019), referred to the increasing interlinking and *mutual* dependency of digital technologies and economies, calling for a move from multilateralism to multi-stakeholderism.

However, the implied mutuality elides the asymmetrical nature of interdependence, and the uneven distribution of rule-making power. Revealing these asymmetries, Global South critical perspectives draw from dependency theory (Ahiakpor, 1985) - the notion that "resources flow from a "periphery" of poor and underdeveloped states to a core of wealthy states, enriching the latter at the expense of the former." As an example, economies and AI

companies of the Global North depend on continued extraction of mineral resources and physical and digital labour from the Global South, and Global South dependencies on the Global North are systematized across consumer goods, digital infrastructure, technology transfer, trade, and financial regulation. The mainstreaming of "interdependence" thus excludes Marxist examinations, critiques of imperialism (Sassoon, 2014; Buecker, 2003) or notions of interdependence as "relational autonomy" from indigenous praxis (Cisneros, 2018; Colloredo-Mansfeld, 2002). These perspectives understand dominant invocations of "interdependence" not as a mutual or symbiotic exchange, but rather a deliberate and continued solidification of exploitative asymmetries in power, information, and natural resource acquisition.

The paradox of participation

"A root cause of failure of developmental projects lies in default attitudes of paternalism, technological solutionism" and predatory inclusion (Mohamed et al., 2020; Taylor, 2019). Inclusion processes in technology policy can often be procedural and numeric, used for virtue signaling and promotional purposes. Furthermore, proportional inclusion can certainly exist while structural harms persist, with "little evidence of the long-term effectiveness of participation in materially improving the conditions of the most vulnerable people or as a strategy for social change" (Cleaver, 1999). This dual reality is described as the "paradox of participation" (Cleaver, 1999; Bliss & Neumann, 2008; Ahmed, 2012), where formal representation can be achieved without any improvement in substantive outcomes, and the distribution of resource, agenda-setting and decision-making power remains status quo (Fraser, 2005).

We often see an expressed desire for global beneficence accruing from the application of AI technologies, across agriculture and healthcare, for example, without action towards mitigating the institutional and systemic reproduction of inequality (Eubanks, 2018). This aspect of apparent benevolence by elite institutions is described by Cusicanqui (2012), Benjamin (2019), Freire (1970), Hattori (2003) and many others. Sylvia Cusicanqui's concept of *gatopardismo* describes the "political philosophy or strategy of advocating for revolutionary changes, but in practice only superficially modifying existing power structures." Ruha Benjamin describes *techno-benevolence*—tech-based interventions that intend to address inequalities, but instead reproduce or deepen dependency and extractivism. Paolo Freire's appraisal of benevolence articulates "in order to have the continued opportunity to express their "generosity," the oppressors must perpetrate injustice as well. An unjust social order is the permanent fount of this generosity... True generosity consists precisely in fighting to destroy the causes which nourish false charity" (Freire, 1970). This "violence of care," coined by Marjo Lindroth and Heidi Sinevaara-Niskanen (2018) vis-à-vis the inclusion of indigenous actors in global climate governance, is often reflected in the logics of inclusion within global governance, AI governance being no exception (Rainie et al., 2019), as well as corporate social responsibility efforts within Big Tech (Nobrega & Varon, 2021; Ricaurte, 2020).

To mitigate the paradoxes of participation, and materially improve substantive outcomes, inclusion efforts within global AI governance would need to both work with Global South stakeholders to develop *formal* regulatory frameworks that address the systemic reproduction

of inequality, and attend to the "wealth of *informal* processes and their politics" where inclusionary trends exist in concert with "more exclusionary tendencies" (Pouliot & Thérien, 2017). Informal processes include hosting global conferences, forming multi-stakeholder coalitions, and mandating groups of experts (Pouliot & Thérien, 2017), as well as informal organizational cultures of coordinating institutions. These cultures reproduce exclusionary tendencies for Global South and civil society actors, and include financial opportunism tied to lobbying and regulatory capture, careerism and political opportunism, protocols of diplomacy which center agendas of powerful actors and exclude oppositional voices, co-option of Global South or civil society terminology and narratives, deficit narratives around developing countries, insufficient technological literacy amongst policy makers writ large, and the use of overly broad policy language that allow for interpretations that protect interests of affluent governments and industry actors (Png, forthcoming).

Interrogating the Global South as peripheral to governance

Before tracing some of the landscape of AI governance discourse by/from the Global South, we must first question the portrayal of Global South actors as peripheral to AI development or governance. Although the integration of Global South stakeholders to the AI governance discourse is still nascent, we need to re-establish and remember the historic agency and existing relevance of the Global South and South–South cooperation in global governance and international norms development. As advanced by Weiss (2016), who challenges the ahistoric characterization of global governance analyses, it is a "convenient narrative," open to critique, that the Global South has had "little impact on universal normative developments," or "was largely absent from the founding of the United Nations whose values came only from the West." This was especially inaccurate for the period of decolonization and post-colonial independence after WWII and into the late 20th century (Weiss & Abdenur, 2014), or during BRICS resistance to OECD agendas (Weiss, 2016). As examined by Helleiner (2014) in "Principles from the Periphery: The Neglected Southern Sources of Global Norms," Global South agency and influence has been "a genuine but essentially ignored source of global norms," from resisting non-beneficial impositions of Western values and blocking asymmetric proposals, to articulating perspectives, policies, and revisionist normative frameworks.

Further, the Global South is neither inoperative, "a passive recipient, nor... the periphery of emerging developments in these data-driven technologies" (Singh, 2021). As elucidated in the following sections, the Global South is very much shaping the development of AI systems: through AI research and industry ecosystems, a source of large-scale commercial data extraction, labour markets for data labelling and annotation, sites for beta-testing of AI systems, and the provision of rare earth minerals and materials required to build physical infrastructures that make AI products and services possible (data centers, graphics processing units, transistors, lithium batteries, etc.) (Parikka, 2015; Crawford, 2021).

It is therefore necessary within inclusive governance efforts to assess structural and agentic framings of the Global South/s within AI governance, in ways that do not fall into the "hype of the rest" (Acharya, 2018) and idealize beyond the realities of long-standing inter and intra-regional inequalities and constraints (political, regulatory, financial, etc.); for example, many low and middle-income countries are still attending to population

access to internet connectivity, electricity grids, or a dearth of required infrastructure for wide-scale implementation of AI systems (Bottino, 2021; Nayebare, 2019; van Stam, 2020).

AI governance and policy discussions must, in light of this, be granular enough to neither impose "tired portrayals" (Birhane, 2020) of deficit narratives, convey unfounded optimism, nor ignore emerging AI research and industry ecosystems emerging out of the Global South, such as tech start-up ecologies in Africa, such as the "Silicon Cape" in Cape Town, "Silicon Savannah" in Nairobi, "Sheba Valley" in Addis Ababa, and "Yabacon Valley" in Lagos (Birhane, 2020). Although AI policy and governance frameworks are disproportionately being debated and developed in Europe and North America—58 percent, in comparison to 1.4 percent in Africa (AI Ethics Guidelines Global Inventory, 2021)—it is important not to sideline activities of different national bodies such as Nigeria's National Agency for Research in Robotics and Artificial Intelligence (NARRAI), Rwanda's Ministry of ICT and Innovation (MINICT) National AI Strategy, or Sierra Leone's Directorate of Science Technology & Innovation (Nayebare, 2019; Romanoff, 2019). Nonetheless, it is yet to be determined how these policy documents will be implemented, enforced, and translated into local or international industry and governmental practices.

A Brief Digest of "AI For and From the Global South"

Drawing from Professor Rafael Capurro's (2008) essay "Information Ethics For and From Africa," which highlights the monopoly of a range of Western ethical traditions in the development of ethical information systems, this section will provide a brief digest of "AI For and From the Global South."

Discourses surrounding the adoption and deployment of AI systems (e.g., digital ID systems) vary significantly across geographies of the geopolitical Global South, operating "on a wide spectrum between optimism of leapfrogging and digital transformation of societies on one end and the pessimism of human suffering caused by new forms of data capitalism and colonialism on the other" (Singh, 2021). Further, regions and countries "have their own specific cultures and infrastructural contexts that will shape what appropriation of data-driven technologies might mean for them" (Singh, 2021).

Over recent decades, we have seen engaged efforts and mobilization towards understanding and centering the needs, interests, values, and influence of the Global South in normative frameworks within fields of information and communications technology (ICT), digital technology, and, more recently, AI. There are emerging scholarly and practitioner communities working to understand the consequences of AI development within political, everyday contexts and histories of the "Global South." There is an ever-widening, transnational discourse of AI ethics and governance constituted by scholars, civil society, and entrepreneurial and government actors concerned with the role of AI systems in the compounding of geopolitical (Couldry & Mejias, 2019) and societal inequalities (Eubanks, 2018). This discourse operates under a different logic than mainstream AI governance spaces, by seeking to foster "forms of counter-hegemonic globalisation" (Escobar, 2004), and emphasizes developing AI regulation and applications which are operational in localized

contexts, or relevant to contextual harms and violations of dignity (Kak, 2020), routinely neglected by empowered decision-makers who work within a dominant status quo.

"Southern" (Connell, 2007) AI discourses tend to engage with the downstream effects of imperial histories, as well as constructive critiques of capitalist structures that scale exploitative, unsustainable, unequal, and harmful practices which remain unquestioned by mainstream AI governance. These draw from areas of postcolonial computing (Irani et al., 2010), decolonial computing (Ali, 2016), data extractivism (Couldry & Mejias, 2019; Ricaurte, 2019; Crawford, 2021), culturally sensitive AI and human rights (Mhalmbi, 2020; Kak, 2020), data colonialism (Birhane, 2020), indigenous data sovereignty (Walter et al., 2020), feminist design practices (d'Ignazio & Klein, 2020), design justice (Costanza-Chock, 2020), and data justice (Milan & Treré, 2019; Taylor 2017). Communities for transnational solidarity and collective action have also rallied to address harmful impacts of AI from the 2017 AI and Inclusion Conference, to work carried out by Article 19, Web Foundation, Research ICT Africa's Africa Just AI Project, Tierra Común, the Non-Aligned Technology Movement, Global Data Justice Project, Technology Justice Lab, Research ICT Africa, Big Data Sur, Black in AI, and Digital Asia Hub, amongst many others.

This body of work is in some ways novel, and in many ways departing from existing areas of critical theory and ICT (Richardson et al., 2006; Fuchs, 2009), intercultural information ethics (Capurro, 2009), critical data studies, new media studies (Wardrip-Fruin & Montfort, 2003), ICT for development (Dey & Ali, 2016), science technology and society, and values in technology (Nissenbaum, 2001; van de Poel & Kroes, 2014; Friedman et al., 2013; Sengers et al., 2005; DiSalvo, 2015).

Critical concerns from the Global South

Digital sovereignty, infrastructural, and regulatory monopolies

Global South adoption of "infrastructural and regulatory landscapes and histories of Euro-America" (Raval et al., 2021) in the form of European Standards Organizations (ESOs), as well as international Standards Development Organizations (SDOs), raise concerns of power consolidation across Western European, North American, and Chinese industries and governments. The production and ownership of technological infrastructure by Global South countries, as well as digital sovereignty (i.e., data ownership, usage, and storage), are essential for countries in the Global South to accrue benefit from AI R&D (Sampath, 2021; Rayment, 1983). These are precursors to participating in the technological advancements and associated benefits advertised by the Fourth Industrial Revolution (Sampath, 2021; Mbembe & Nuttal, 2004).

Nonetheless, issues outlined above, as well as first mover advantages in trade, unequal public–private partnerships, models of manufacturing, procurement protocols, cost of development, and pricing of technologies have not yet been identified as sites for necessary reform in mainstream AI governance discourse, requisite to assuring wider international benefit of AI technologies.

Sovereignty

Kovacs and Ranganathan (2019) caution against any uncritical operationalization of sovereignty, reminding us that "it is important to ask under what conditions it becomes possible

to reclaim sovereignty despite these violent roots." The justification of territorial and digital sovereignty, and concomitant self-determination differ greatly between systems of governance (e.g., European, African, or Indigenous systems of governance), domestic and foreign policy agendas, stages of technology infrastructure development, particularities of national regulation, and IP legislation. Given this variability of notions of sovereignty, for example, tensions between state and Indigenous sovereignty (Bauder & Mueller, 2021), understanding which are legitimized in AI governance discourse and which are sidelined, and why, is crucial. For example, the Indigenous Data Sovereignty movement, advocating for the "right of Indigenous peoples to control data from and about their communities and lands, articulating both individual and collective rights to data access and to privacy" (Rainie et al., 2019), embody valuable examples of beneficial governance, but present legitimate challenges to state actors leading the dominant AI governance discourse.

Pricing and ownership

As described by Sampath (2021), the pricing of exported products "is not determined by a mark-up price set by the Southern producer, but continues to be dependent on the demand generated in Northern export markets." This precludes developing countries' self-determined industrialization and development, leading to the continuation of structures that encourage extractive dependencies (Grosfoguel, 2016) of Global South countries on Global North private companies and regulators. Financial monopolies of rich industrialized governments and industries, as well as political power and trade advantage, are characteristic of modern-day globalized capitalism (Wade, 2013; Karatasli, 2020; Piketty, 2014).

Similar monopolies and dependencies are observable in digital infrastructure. GAFA companies (Google, Amazon, Facebook, and Apple) hold high market share in Global South countries, while maintaining infrastructural monopolies that create extractive technological dependencies (Kwet, 2019). When analyzing GAFA companies, Rosa and Hauge (2020) find that their points of interconnection, essential for accessing content delivery networks (CDNs), are concentrated in Global North countries, with more than twice as many countries in the Global South having no points of interconnection. This means that internet service providers (ISPs) in the Global South incur higher access costs, while revenue-generating data move unidirectionally from the South to the North (Rosa & Hauge, 2020); this is described as economic domination (Kwet, 2019). Given this asymmetry in financial monopolies and infrastructural dependencies between the Global South and North, Rosa and Hauge (2020) argue that governance debates "should be reconfigured to account for the complexities of privatized digital information infrastructure and the extraterritorial effects of U.S. laws embedded into the design of these platforms."

The unfettered access and monetization of Global South national data by Big Tech is detrimental to growing local data economies. In a similar account of unequal infrastructural ownership, African critical data infrastructure (submarine cables, terrestrial fiber-optic networks and data centers) are majority-owned by non-African telecom companies. Given that many African countries do not have national data centers, sensitive population data are largely hosted on servers abroad (e.g., in Ireland). Studying the dynamic between South Africa and the United States, Kwet (2019) asserts "foreign powers, led by the US, are planting infrastructure in the Global South engineered for their own needs... imposing privatised forms of governance... This allows them to accumulate profits from revenues

derived from rent (in the form of intellectual property or access to infrastructure) and surveillance (in the form of Big Data)." In response, initiatives such as Smart Africa and the African Tax Administration Forum are developing privacy and taxation policies for Big Tech (Velluet, 2021; Elmi, 2020) germane to AI governance discussions.

Intellectual property

Monopolies of AI-relevant IP and regulatory power by rich industrialized governments and proprietary technologies and software held by Big Tech further preclude the autonomy of Global South countries and their ability to accumulate profits from their own data (South Centre, 2020; Kwet, 2019). This is despite a significant amount of technological innovation being driven by "public institutions or with a heavy dose of public funding"; for example, the internet, GPS, the Siri virtual personal assistant, Google's search algorithm, and lithium-ion batteries (Kwet, 2019; Mazzucato, 2018).

Current solutions of technology transfer (Gopalakrishnan, Shanthi, and Michael D. Santoro, 2004) and technical assistance from North to South are also impeded by patent rights, which according to Intellectual Property Watch (Cheikh, 2010), can "severely reduce technology transfer since they bring high licensing fees and can thus impede the knowledge adaptation to local conditions." IP agreements such as the Trade-Related Aspects of Intellectual Property Rights (TRIPS) are adopted by developing countries as a membership requirement for the WTO, but have been contested for their asymmetric benefits; in 2004, developing countries proposed an Intellectual Property agenda "that would bridge the 'knowledge gap' and 'digital divide' through 'public interest flexibilities' " (WIPO, 2004). Similar contestations are currently underway for critical COVID-19 supplies, where countries such as the United States, the United Kingdom, and certain EU countries have blocked the global scaling of vaccine production, prioritizing pharmaceutical profits through the obstruction of the TRIPS waiver at the WTO, which would suspend patents on certain COVID-19 medical tools (Vawda, 2021; Working Group on Trade Treaties and Access to Medicines, 2020). Rectifications to such monopolies are highly complex, but suggested solutions include guidelines development within regional IP offices based on South–South cooperation, and reforming private sector incentives.

Extractive logics as a systemic pattern

"Extractive logics as a systemic pattern" of data economies are increasingly being accounted for (Jameson, 2019)—from large-scale extraction of information by capitalist enterprises (Couldry & Mejias, 2019; Ricaurte, 2019), to surveillance capitalism (Zuboff, 2019), to global surveillance capitalism wherein "Global North intelligence agencies partner with their own corporations to conduct mass and targeted surveillance in the Global South" (Kwet, 2019), all the way to the extraction of rare earth minerals used to build physical infrastructures underpinning AI systems.

Pollicy, a Ugandan research institute, map "digital extractivism" from cheap digital labour, to "illicit financial flows, data extraction, infrastructure monopolies, digital lending, funding structures, beta testing and platform governance" (Iyer et al., 2021; Teevan, 2021). Sampath (2021) describes extraction from low- and middle-income countries by the technology industry, through introducing "as many applications, platforms and other

digital products/services as possible in order to extract the maximum amount of data." This represents "a shift from valuing people as consumers to extracting value through a proliferation of complex instruments" (Sassen, 2017).

In a coherent articulation, Paola Ricaurte (2019) explains that, "Data-centered economies foster extractive models of resource exploitation, the violation of human rights, cultural exclusion, and ecocide. Data extractivism assumes that everything is a data source. In this view, life itself is nothing more than a continuous flow of data."

Material extractionism

Environmental concerns are still peripheral within the dominant AI governance discourse, and although emerging (e.g., GPAI's "A responsible AI Strategy for the Environment" [Clutton-Brock et al., 2021]), most work engaging on environmental sustainability, such as Green AI or AI & Sustainability overlook concerns raised by Global South advocates. While AI can indeed be used to optimize energy use and support renewable energy (e.g., green technologies), the argument that AI is a solution to the climate crisis is circuitous, and governance frameworks cannot omit the environmental costs of AI/information infrastructure (Parikka, 2015). As Crawford (2021) reiterates: "The data economy is premised on maintaining environmental ignorance."

Contrary to a conventional focus on the immaterial nature of AI, Global South-centered AI discourses call to attention the obfuscated human labour (Gray & Suri, 2019) and material infrastructural components (Parikka, 2015) of AI, in order to underscore the wider impacts AI systems can have across human and environmental registers. Tapia and Peña (2020) articulate that "dominated by a liberal framework, the material conditions of production of technological devices that allow digital communications are still ignored"—the "Weight of the Cloud," so to speak (de Vicente, 2013).

Discerning the responsibility of AI companies regarding environmental sustainability and developing avenues for accountability is of critical importance. The energy cost of training machine learning models or developing natural language processing systems is increasingly incorporated into mainstream models of AI harms (Bender et al., 2021; Strubell et al., 2019). Nonetheless, still neglected are harms from intensive water and fuel usage of server farms (Mytton, 2021), consequent chemical and e-waste, as well as the opaque supply chains of tech/AI company physical infrastructures, described by Abraham (2017) as "a murky network of traders, processors, and component manufacturers... the hidden link that helps in navigating the network between metals plants and the components in our laptops."

AI companies engage in geological extraction through investments in smelters and mining concessions for mineral sourcing needed for hardware. In addition, contracts by Microsoft, Google, and Amazon have provided tools to the oil and gas industry for extraction optimization (Greenpeace, 2020), further degrading our environment and ecological services. The buds of such considerations are appearing dominant AI policy discourse (Brevini, 2022). In a European Parliament report, Bird et al. (2020) recognize that "The extraction of nickel, cobalt and graphite for use in lithium ion batteries—commonly found in electrical cars and smartphones—has already damaged the environment, and AI will likely increase this demand" (Khakurel et al., 2018). This analysis also requires that we look at the human costs of environmental degradation, which are disproportionately carried by the Global South. Extractive industries similarly first impact racialized, vulnerable,

and neglected groups via exploitation, state and mining industry violence on indigenous communities, and increased gender violence (Legassick, 1974.

Data labour markets and workers' rights

There is a growing movement and literature surrounding the role of workers in AI governance (Nedzhvetskaya & Tan, 2021). A whole-systems assessment of risks and costs of AI systems recognizes the underlying human labour. This includes e-waste workers, including children, who work on digital dumpsites and are exposed to e-waste toxicants that are also part of AI infrastructure lifecycles (WHO, 2021; Browne, 2021; Ohajinwa et al., 2017; ILO, 2019). "Ghost workers" (Gray & Suri, 2019) are low-paid workers who drive the AI economy—they annotate and classify large volumes of data to improve computer vision, natural language processing, or other types of algorithms. These workers are often hired from countries in the Global South, optimizing for labour costs, contracted by specialized annotation platforms such as Microworkers, Samasource, CrowdFlower, or Amazon Mechanical Turk. These companies indeed provide jobs, but they lack appropriate policies mitigating workers from exploitative industry practices. Without accountability structures, employers have been known to withhold remuneration, denying "the rights of workers to safer, dignified working conditions" (Irani & Silberman, 2013), disproportionately impacting those who are economically vulnerable, especially acute in countries with limited labour protection laws (Yuan, 2018).

Tech worker coalitions such as Turkoptikon, labour unions such as UNI Global Union, or research initiatives such as Fairwork are essential in shaping industry and policy level practice, providing a worker-centered understanding of risks and harms. "Unskilled laborers" are excluded from "expert"-centered governance discussions on the automation of work, and are nonetheless both essential for the development of AI systems as well as robust protective regulations.

Beta-testing, commercial exploitation, and contextual incompatibilities

Beta-testing is a form of commercial exploitation, entailing "the testing and fine-tuning of early versions of software systems to help identify issues in their usage in settings with real users and use cases" (Mohamed et al., 2020). This practice exhibits a pattern of selecting populations that are systemically more vulnerable to risks, or jurisdictions that lack pre-existing safeguards and regulations around data usage, given this practice would violate laws in their localities (UNCTAD, 2013). For example, predictive policing software developed by Palantir and used by the New Orleans Police Department disproportionately on the African American community, or the use of election analytics in the Kenyan and Nigerian elections by Cambridge Analytica before their deployment in Western democracies (Mohamed et al., 2020). It is therefore crucial to resource capacity for low- and middle-income countries to strengthen legal and institutional protections of marginalized people's rights in ways that work to address long-standing exploitative commercial practices.

Another form of misalignment between product providers and users is the contextual incompatibility of AI systems resulting from a lack of developing country-specific training datasets (Quinn et al., 2014), or misrepresentation of populations and their socio-cultural

or political behaviour in existing training datasets. For example, a learning algorithm trained on North American traffic flow or public health datasets is unable to be directly implemented in Central America, Africa, or Asia, risking misapplication (Neupane & Smith, 2017). Global South discourses are grappling with geographic core-periphery patterns of data availability (Graham, 2014) and asking who is able to collect, access, or be represented—and for whose gain (Graham et al., 2015).

Key Differences Between Dominant and Global South discourses: What Does This Mean for AI Governance?

"Minimizing risks and maximizing benefits" has been an important directive in orienting the AI ethics, safety, and governance discourse in the Global North. The identification of immediate and tail-risk harms at different scales range from runaway AI (Kurzweil, 2005) and lethal autonomous weapons to devastating de-democratizing effects of AI driven platforms, and areas of "bias and fairness, accountability, transparency, explainable AI, and responsible AI" (Singh, 2021).

Certain Global South discourses expand these concerns to wider systemic issues outlined above as critical concerns, which understand AI R&D and industry within a geopolitical and historic context of exploitation between the North and South by way of infrastructure, trade, regulation, geopolitical and financial power, industry monopolies, epistemic hierarchies, extraction of natural resources, etc. The absence of such issues from dominant AI governance discourse makes it clear that involvement of Global South constituents—civil society, academia, companies, and governments—is necessary for more comprehensive risk assessment and governance of AI systems.

Beyond Status Quo AI Governance: Roles of the Global South in Global AI Governance

As Chan et al. (2021) argue, "more focus should be placed on the redistribution of power, rather than just on including underrepresented groups" in order to mitigate a future where AI systems "are unsuited to their conditions of application, and exacerbate inequality." The purpose of inclusion is therefore institutional reform—the redistributing resource allocation, agenda-setting and decision-making power (Fraser, 2005). Based on the spectrum in inclusion–exclusion dynamics provided by Marchetti (2016) of ostracization, exclusion, co-option, inclusion, and integration—to which critical views would add structural reform, dismantling, and alterity—inclusion is only the first positive step away from exclusion.

Within this view, internal efforts towards inclusion are necessary, but primacy is conferred to contesting and interrogating the very geopolitical, economic and governance structures that Global South actors are being included into. Postcolonial, or "Southern"

views (Comaroff & Comaroff, 2012) challenge established AI governance at multiple levels, such as which issues are conferred legitimacy, unquestioned ideological assumptions within governance processes, discursive power and its financial drivers, regulatory power, etc.

Efforts towards meaningful inclusion, and critically led integration of Global South actors into AI governance processes requires an intentional formalization of their roles. These roles must be co-constructed with Global South actors—from civil society, industry, academic institutions, and governments—alongside a negotiation to reform exclusionary dynamics in governance processes. Proposed roles of Global South actors proposed by this chapter are three-fold: (1) acting as a challenging function to exclusionary governance mechanisms; (2) providing legitimate expertise in the interpretation and localization of risks, demands, and issue-framing, which includes providing a whole-systems (e.g., supply chain to deployment) and historic view of AI systems; and (3) providing a source of alternative governance mechanisms; e.g., South–South solidarity, co-governance, democratic accountability, and a political economy of resistance.

Challenging function to exclusionary governance mechanisms

A key role of the "Global South" within global governance systems, including AI governance, is fundamentally a challenging function to dominant legal and governance structures that disproportionately support the accretion of safety and capital to certain groups and countries, to the neglect of those marginalized (Marchetti, 2016). Global South state and non-state actors have long intervened upon, challenged, and shaped international law, trade agreements, intellectual property law, and protocols of technology procurement from commercial providers. Non-state actors provide the body of a "political economy of resistance"—a subset of which has been described as "data activism" (Taylor, 2017; Torres, 2017; Milan & Trere, 2019)—ranging from back-engineering statistical models (Pasquinelli & Joler, 2020) and advocacy for facial recognition moratoriums to applying strategic litigation and co-developing policy proposals with directly affected communities. These practices articulate "new forms of political participation and civil engagement in the age of datafication" (Milan & van der Velden, 2016).

It is, however, reductionist to assume that all Global South state actors adopt anti-hegemonic or postcolonial views. Indeed, many "emerging economies appear to have preferred the status quo and working within existing institutions created by Western states." This may however change as middle income countries continue to accrue economic power, "as they grow in power and seek to ensure that their needs and values are reflected at the global level, their assertiveness and dissatisfaction with existing institutions may rise" (Held & Rogers, 2013).

Further, efforts for equitable distribution of benefits between countries at an international level (trade negotiations, procurement contracts, standards setting, intellectual property rights, etc.) do not resolve domestic issues of inequality and harm. State repression and structural violence supported by government action exist across countries globally, with civil society acting as a challenging function. As Global South actors challenge exclusionary AI governance mechanisms, and endeavor to shape AI governance processes, tensions between state and non-state actors must also be a point of strategic focus because government level and citizen level goals often diverge. For example, the Kenyan government's National

Integrated Identity Management System was contested by the Nubian Rights Forum and the Kenya Human Rights Commission who took legal action on the basis that the system violated "the right to privacy, equality, and non-discrimination enshrined in Kenya's constitution" (Mahmoud, 2019).

Legitimate interpretation of harms

"The key feature of transnational activism in global governance is precisely its stubborn attempt to influence the normative battle on the right and legitimate interpretation of crucial global issues" (Marchetti, 2016). The aforementioned critical neglected areas—digital sovereignty in light of colonial histories, infrastructural and regulatory monopolies, material supply chains of AI systems, data extractionism, beta-testing, commercial exploitation, regulatory capture, and workers' rights—are current in Global South discourses but not yet conferred sufficient legitimacy to be integrated in global AI governance agendas.

The integration of these issue spaces should be steered by those who experience the costs incurred by AI systems and their wider political economies. Governance decision-making power is held by elite groups constrained largely to Europe and North America and composed of bureaucrats, industry leaders, and regulators who are often removed in their experience and knowledge from the realities of how harms operate (Mcglinchey et al., 2017). It is in fact those at the sites of harm who practically understand required solutions.

Efforts to govern and regulate the building and deployment of AI systems, especially in the development of protective guardrails, often neglect the legitimate capacity of Global South state and non-state actors to interpret the harms which they are subject to. Regional, financial, and institutional credibility results in elite institutions, industry, and governments—largely based in Western Europe, North America, and China—are disproportionately conferred legitimacy in defining AI-mediated harms and their corresponding mitigating actions. Civil society and Global South actors are conferred disproportionately less legitimacy, visibility, and influence. This pattern within international governmental organizations surfaces a double standard, wherein they simultaneously rely on civil society to legitimize the democratic quality of their governance activities, whilst undermining their legitimacy (Dryzek, 2012).

Recognizing the legitimate expertise of impacted groups counterbalances the tendency of AI governance initiatives to assume universal notions of harm, where definitions are inaccurately generalized in ways that neglect cultural, regional, and jurisdictional divergences. Assumptions of universalisms often entail hegemonic impositions of ideologies or blueprints of progress particular to a powerful minority. In the case of AI governance, governments of information mature economies and industry in North America and Western Europe (Mignolo, 2012; Gramsci, 1971). This is highlighted by Arora (2018) within digital privacy "as technology companies expand their reach worldwide, the notion of privacy continues to be viewed through an ethnocentric lens. It disproportionately draws from empirical evidence on Western-based, white, and middle-class demographics." "Southern" philosophical perspectives of privacy, definitions of privacy-related harm, use-cases of what privacy protection looks like, and locally germane legislation—where privacy regulation "dignifies those at the margins, by giving their privacy its contextual integrity"—are therefore essential. For example, the South African government's Smart ID in the context of South Africa's history,

evokes for certain citizens "past identification systems such as the Population Registration Act by the apartheid government used to racially segregate citizens" (Arora, 2018). In order for definitions of harm and regulatory standards to be contextually operationalizable, comparable rather than universal global standards have been proposed as part of the coordination of global regulatory responses to AI technologies (Abdala et al., 2020). The proposed ideal is interregional coordination of global standards whose localization is supported by participatory methodologies from the margins (McDowell, 2016).

A whole-system view of harms

As discussed earlier, an emerging role of Global South actors in AI governance is to broaden the assessment of AI risks (Kak, 2020; Garcia, 2021) to include risks not only from AI deployment, but also from their material production. Supply chains traversing the Global South are often rendered invisible to consumers and policy makers in the Global North by capitalist production networks (Star, 1990; Star & Strauss, 1999) and immaterial narratives of AI: "ignorance of supply chains is baked into capitalism" (Crawford, 2021; Star, 1990).

Harms resulting from profit models which exploit cheap labour and natural resources from the Global South, and unequal recourse available to exploited groups—who account for a global majority—are more visible at the peripheries of capitalism. The AI industry's expanding supply chains—including mineral suppliers and smelters—impact on the rights and health of people, and the vitality of ecosystems, reflected in long-standing contestations surrounding physical–digital infrastructures led by communities impacted by the extraction industry (GISWatch, 2020). This broadening provides a more comprehensive and whole-systems view of harms, which traverses AI industry supply chains all the way to the implementation of systems in different sectors of society—the life cycle of AI (AHEG, 2020). This "Southern" view contrasts with the dominant conceptualization of AI as artificial, immaterial, decoupled from geologies, ecologies, and people—understanding AI a product of an environmentally costly supply chain, and reliant on a growing labour market of workers.

A historic assessment of harms

Historic perspectives are often sidelined by futurist technological idealism. Given the nascency of AI governance, it is crucial that historic and systemic inequalities replicated in the design, deployment, and regulation of AI technologies are mitigated early on. Within AI governance, there is the risk of "locking in" (Garfinkel, 2018) or compounding historic inequalities in ways that are difficult to mitigate later on. Locking in (Garfinkel, 2018) describes the risk that "certain decisions regarding the governance or design of AI systems may permanently lock in, in a way that propagates forward into the future in a lastingly positive or negative way." Certain Global South discourses affirm that what is projected forward is carries forth cumulative histories, which have shaped our present dependency dynamics between the Global South and North, legal systems, physical and digital infrastructures, and social hierarchies "that continue from the past and remain unquestioned in the present" (Mohamed et al., 2020).

Providing alternative governance mechanisms

> *We need to change the institutions that have historically been set up as tools of advancement and control for some, to the exclusion of many, and not just tweak it or look for ways to create space. Without this, advancement of science and technology will continue to benefit those who governed historically at the expense of those who were excluded.*
>
> (Sampath, 2021)

As aforementioned, one of the roles of Global South actors within AI governance is a challenging function to the inequitable distribution of the benefits and costs of AI systems, and interrogating the governance, geopolitical, and economic infra/superstructures that AI industries and their regulations that shape this inequality.

There is a distinct community of Global South-oriented practitioners developing new forms of political objectives, praxis, including action repertoires and strategic organization (Milan & van der Velden, 2016), that describe alternative mechanisms for governance. Some of these concepts and practices of reform are summarized in the following section, structured under Stefania Milan's (2013) framework outlined in "Social Movements and Their Technologies," classifying different action repertoires of mobilization as "inside," "outside," and "beyond" institutions of decision-making power. These include participation, co-governance, and mechanisms for democratic accountability within institutions of power, efforts of South–South solidarity oriented outside institutions of power, and a generative political economy of resistance oriented beyond institutions of power.

Inside: Participation, co-governance, and democratic accountability

Participation and co-governance

Co-governance is a more developed framework of inclusive governance, seeking not only to achieve proportional representation but to also legitimize interpretations of harms and "shift the balance of power in favour of crisis-affected communities" (Berditchevskaia et al., 2021). It understands that AI risks cannot be adequately defined by those who are distanced from these risks by dimensions of power and institutional safety (Ulnicane et al., 2021; Milan & Trere, 2019; Schiff et al., 2021), and that legitimate assessment of harms and building of guardrails is not possible without participation of Global South state and non-state actors. As Kak (2020) articulates, there is a "need to approach the political economy of AI from varying altitudes- global, national, and from the perspective of communities whose lives and livelihoods are most directly impacted in this economy." The slogan "nothing about us without us"—originating from Central European political traditions (Smogorzewski, 1938) and later adopted in the disability rights movement around the development of innovative technologies (Werner & PROJIMO, 1998)—emphasizes that no policy should be developed without direct participation of affected communities.

Centering the knowledge of those most exposed to risks has been long developed by Participatory Action Research, and Critical Development Studies (McDowell, 2016). Participatory artificial intelligence entails the "involvement of a wider range of stakeholders than just technology developers in the creation of an AI system, model, tool or application"

(Kulynych et al., 2020); for example, product risk assessments carried out with impacted communities, recognizing the expertise of those who experience the costs incurred by AI systems (Frey et al., 2018). This is to ensure that AI technologies, and associated protective regulation, are able to contextually respond to "specific social, economic, and cultural demands" (Abdala et al., 2020).

Participatory practices (Gangadharan, 2013) are more common within NGO and public sector entities, which are "generated through processes that are more participatory" (Schiff et al., 2021) than private sector or international organizations, have "more ethical breadth in the number of topics covered, are more engaged with law and regulation." The ethical breadth and contextual specificity of NGO and public sector entities is lacking in policies developed by private entities or international organizations, instead using broad language that indirectly allows for interpretation that conveniently protects interests of industry actors and wealthier states. For example, the high-level expert group on artificial intelligence's (AI HLEG) draft laws for AI regulation, one of the most significant attempts at AI regulation, was critiqued by civil society who noted the broad nature of draft laws left a range of discretion for industry actors to self-regulate by stating that "many industry groups expressed relief the regulations were not more stringent, while civil society groups said they should have gone further" (Kind, in Satariano, 2021).

"Democratic accountability"

It is well understood that there are accountability deficits in global governance bodies "which lack formal mechanisms of democratic accountability that are found in states, such as popularly elected leaders, parliamentary oversight, and non-partisan courts. Instead, the executive councils of global regulatory bodies are mainly composed of bureaucrats who are far removed from the situations that are directly affected by the decisions they take" (Dryzek, 2012). Recognizing healthy critiques of NGO accountability to communities (Charnovitz, 2006), McGlinchey et al. (2017) offer, "civil society action at the international level is predominantly focused on building political frameworks with embedded democratic accountability." The concept of "counter-democracy" (Rosanvallon & Goldhammer, 2008) is useful here, not contrary to democracy, but "a vital and perennial aspect of it... counter-democratic actors organise distrust against power-holders, pressuring them to strengthen accountability" (Kalm et al., 2019).

Co-governance for accountability has been designed into program in Brazil, Mexico, the United States, and India across policy areas ranging from participatory budgeting, anti-corruption, poverty reduction, infrastructure provision, school reform, electoral administration, and police reform (Ackerman, 2004). Some of these programs have achieved significant pro-accountability success by "giving social actors direct access to state institutions," and creating conditions where citizens can challenge governments through the media or courts (Ackerman, 2004). This form of vertical accountability is critical to AI governance, as seen in the series of facial recognition moratoriums led by civil society such as the American Civil Liberties Union (ACLU) in coalition with impacted communities, as well as contestations of AI industry regulatory capture (Whittaker, 2021; Saltelli et al., 2022; Ochigame, 2019; Metzinger, 2019). Although industry is essential for functional AI governance, as a "nongovernmental and noncommercial space of association and communication" (Jaeger, 2007) civil society is a mechanism for people-centered accountability, beyond

profit incentives which have transgressed fundamental rights. Civil society organizations operate by participating in policy reviews, increasing public visibility and transparency of global governance processes, working towards redress of harmful externalities from regulation or lack thereof, and advancing formal accountability or enforcement mechanisms such as strategic litigation for governance operations (Scholte, 2004). This work benefits all sectors of society, and is often under-resourced and under-compensated, and with global AI initiatives often requesting civil society to participate in consultations on a voluntary basis McGlinchey et al. (2017)—highlights the need for such initiatives to match these invitations with resources and compensation.

Outside: South–South solidarity

South–South coalition building and solidarity is an important asset for low- and middle-income country governments to multilaterally advance advantageous agendas within AI governance fora that are often not incentivized to support these agendas. Geopolitical coalitions such as the UN South–South Initiative, Group of 77, and the Non Aligned Movement—key in the decolonization and independence movements in Africa, Asia, Latin America, Middle East, and other regions—are examples of alternative governance mechanisms that can be engaged with through AI global governance initiatives. Today, these coalitions represent two-thirds of UN membership, and 55 percent of the global population.

Although they are state-led, grappling with aforementioned state–civil society tensions, they present opportune mechanisms for Global South governments to assert self-determination in AI governance processes through multilateral action, develop and uphold collective and anticipatory interests, and engage in capacity-building in ways that interrupt extractive dependencies on Global North governments and technology firms (Weiss, 2016). Importantly, these coalitions embody an experiential and institutional memory that reaffirms and articulates continuities of colonialism in contemporary global inequality.

In 2020, Professor Ulises Mejias proposed a Non Aligned Technology Movement, with a primary goal of transitioning from technologies that reinforce aforementioned dependency dynamics, to technologies that support the self-determination of developing countries (Mejias, 2020). The design, development, and deployment of AI technologies, and their governance, are being further re-thought by indigenous-led groups such as the Global Indigenous Data Alliance, and others aligned with postcolonial self-determination. These calls for reformist or "alternative" governance mechanisms comprehend how bureaucratic, cultural, economic, and epistemological legacies of European colonialism are integrated within global governance structures emergent out of the postcolonial era (Quijano, 2000; Sampath, 2021). The apprehension of these histories is a prerequisite for meaningful Global South participation, robust discussions of risk mitigation, and preventing locked-in inequalities.

Beyond: A generative political economy of resistance

Milan and Treré (2019) conceptualize the Global South, as a "place of (and a proxy for) alterity, resistance, subversion, and creativity," generative of alternative model blueprints. Questions as to how AI systems should be designed, developed, deployed, and governed are being asked by

groups (some previously discussed) beyond government and industry influence, centering the day-to-day realities and volition of communities undermined by institutional and commercial practices. Resistance practices which seed the possibility of alternative designs and use, or non-use, of AI systems describe a global "political economy of resistance" (Taylor, 2017). These include instances of algorithmic resistance in Latin America (Treré, 2018), counter-surveillance (Gürses et al., 2016), grassroots data activism (Milan & Gutiérrez, 2015), and new data epistemologies (Milan & van der Velden, 2016). Anti-possessive frameworks include the Free Software Movement (Stallman, 2002), which developed "publicly owned and controlled technology built for freedom by design at the architectural level" (Kwet, 2019), and free/libre and open source software (FLOSS) endorsed for public sector implementation across the Global South (Papin-Ramcharan, 2007). Alternative systems and hardware solutions are also being built at a local level—from "creole technologies" (Edgerton, 2007) to postcolonial computing designs (Irani et al., 2010), to innovation ecosystems based on collection and ownership of data by local communities (Ricaurte, 2019).

Power Analysis: Explaining Key Differences

> *A defining property of a socio-technical regime is the interdependent, highly nstitutionalized alignments across heterogeneous processes that serve to reproduce the regime, and which tend to engender path-dependent development. This constitutes a form of structural power which privileges certain actors at the expense of others.*
>
> (Smith & Stirling, 2008)

Recognizing the incongruities between Global South and Global North analyses of AI risks and understanding that more than just being passively represented, Global South actors can engage in productive institutional change by interrogating AI governance processes, are essential undertakings for those seeking to ensure that AI governance can materially mitigate risks and equitably distribute benefits for a global majority.

More fundamentally, this understanding must be embedded within a "broader analysis of power and political dynamics" (Ó hÉigeartaigh et al., 2020; Chan et al., 2021; Kalluri, 2020). Global North geopolitical, industry, and regulatory dominance are outcomes of heterogeneous processes that reproduce regimes of structural power, privileging Global North actors at the expense of Global South actors (Smith & Stirling, 2008). These processes occur within "institutions that have historically been set up as tools of advancement and control for some, to the exclusion of many" and governed in ways that support this (Sampath, 2021). Not only are mainstream AI governance processes developed and operationalized within such institutions, these institutions exhibit a gap in understanding or acknowledging historic processes that result in the underrepresentation of Global South actors and reproduce the consolidation of resource power, and regimes of structural inequalities.

The coloniality of power in AI governance

AI governance is particularly prone to gaps in understanding or acknowledging historic processes because ideals of technological progress and industrialization, looking to the

future, often obfuscate our histories. Such ideals are "deeply entrenched in a wider social context that encourages us to ignore the historical roots of current inequalities—which, in fact, are not amenable to a technological solution alone" (Sampath, 2021). More fundamentally, governance suffers from a "historical amnesia," deducing "general laws from the specific historical context during which they emerged, while at the same time invisibilizing this temporal embedding" (Albert & Werron, 2020).

We cannot understand present AI harms, or anticipate their futures, without understanding their historic trajectories. Exclusionary geopolitical path dependencies, economic first-mover advantages replicated by the AI industry (Pathways for Prosperity Commission, 2019), as well as the automation of inequalities (Eubanks, 2018) have important material roots in European colonialism (Mohamed et al., 2020). Nonetheless, global governance omits "the imperial and (post)colonial past and present of international relations, thereby presenting a theoretical picture of world politics that is deeply embedded in Eurocentrism and therefore exhibits serious theoretical and empirical flaws" (Albert et al., 2020; Hobson, 2012).

Contemporary legacies of European colonialism in interpersonal, societal, and geopolitical power asymmetries are described by the concept of "coloniality" (Quijano, 2000). Coloniality is the continuation of inequality along dimensions of authority, economy, politics, race, gender and sexuality, knowledge, and subjectivity (Maldonado-Torres, 2007), emergent from "the historical processes of dispossession, enslavement, appropriation and extraction... central to the emergence of the modern world" (Bhambra et al., 2018). The coloniality of power informs emerging frameworks—such as data colonialism and data capitalism—that recognize continuities of colonial exploitation, extraction, and dispossession in the Global South, in the use of labour, material resources, and data in AI lifecycles (Thatcher et al., 2016; Ricaurte, 2019; Couldry & Mejias, 2019; Birhane, 2020; Zuboff, 2019; Irani et al., 2010; Ali, 2016).

Within this framing, we can understand the perhaps genuine intention to develop inclusive governance as missing a core macro-historic analysis of how institutions, governments, and bureaucracies shaping AI governance participate in legacies that reproduce structural injustices in latent or active ways. For example, international laws which would underpin the global governance of AI systems are evidenced to have been constructed by and rooted in institutions of geopolitical domination and imperialism. As Raval, Kak, and Calcaño (2021) argue, "in post-colonial societies governance infrastructures and legal frameworks are also shaped by colonial legacies" resulting in "problematic efforts at digitalization and complex dependencies on foreign enterprise" and as such "unreliable, dynamic and highly contested practices of data governance." There is therefore a need for AI governance to continuously question the histories of legal systems and what harms they preclude and enable, as Sampath (2021) elucidates, "what has historically been considered to be legal, and institutionalised through formal rules, is not necessarily moral or desirable for an equitable future."

Tools for interrogation can be drawn from critical legal scholars such as Anthonie Anghie, associated with the Third World Approaches to International Law (TWAIL) movement, which integrates the historic colonial context of international relations (Anghie & Chimni, 2003). This is especially pertinent to critical discourses of global justice within AI governance, which articulate the limitations of Western ethical values (IEEE, 2019; Raval et al., 2021; Arora, 2019; Gabriel, 2020), which are "too parochial and eurocentric to meet

global challenges" and must engage with complex but necessary pluralism across "Hindu law, Muslim law, African laws and Chinese law" (Menski, 2006).

This is not to dismiss the merit and efficacy of international law, but rather to highlight that colonial legacies and eurocentrisms embedded within have not been efficacious in benefiting the autonomous interests of Global South actors. As Chimni (2006) articulates, "the relationship between State and international law is being reconstituted in the era of globalization to the distinct disadvantage of third world States and peoples." This balance between critique in order to improve, and the practicality of available tools, is apparent within discussions of Human Rights, and its applicability to mitigating AI harms (Smuha, 2020, Risse, 2018, Mantelero, 2018). Scholars such as Samson (2020) note that the universality enshrined in Human Rights is not available to everyone, historically "embedded in deep, exploitative structures... hand in hand with the systematic privileging of some groups over others," where the humanity of individuals, and thus their entitlement to protection, is subject to different constructed hierarchies, such as racism (Robinson, 1983). This can be held as an accurate appraisal, whilst using Human Rights as effective safeguarding tools, "enabling legal obligations that are already enforceable today, also in the context of AI" (Smuha, 2020) and taking the opportunity to assess both "the applicability and vulnerability of human rights in the context of AI" and the "current enforcement mechanisms of these rights" (Smuha, 2020).

Historicizing popular AI governance discourses: The Fourth Industrial Revolution

Under the Fourth Industrial Revolution (4IR), the AI industry is estimated to produce an "additional economic output of around US$13 trillion by 2030, increasing global GDP by about 1.2 % annually" (Bughin et al., 2018). Although economic benefits may indeed accrue to certain Global South economies and across healthcare, communications, agriculture, jobs, and education, these claims do not recognize longstanding exclusionary path dependencies and first-mover advantages within commercial trade, or infrastructural initiatives and partnerships. As Chan et al. (2021) articulate, "those best-positioned to profit from the proliferation of artificial intelligence (AI) systems are those with the most economic power," corroborated by Murphy et al. (2014) who conclude that "ICT integration is, on balance, benefiting foreign-owned businesses and corporations."

Economic benefits of AI often fall under "hyperbolic claims that big data and the data economy are the new 'frontier of innovation,' with 'cost-effective,' 'profit-generating' properties for all" (Sampath, 2021), and do not recognize selective enrichment and market concentrations that the AI industry enables, resulting in a "Matthew effect" (Wade, 2004; Fernández-Villaverde et al., 2021), which deepens inequalities between the Global North and South as historic inequalities "intertwine with new power asymmetries to create newer, and more drastic, degrees of exclusion" (Sampath, 2021).

It is important to note that both notions of Industrial Revolution and Global South are closely tied to colonial histories, and although these terms are used regularly by the dominant AI governance discourse, their associated histories are omitted, in part because their impacts do not overtly damage rich industrialized countries. Present-day exclusionary

path dependencies are congruent with prior industrial revolutions—the First Industrial Revolution relied on the dispossession of land, brutal exploitation of people's labour, extraction of knowledge and natural resources from European colonies, and strategic underdevelopment—made possible by military-led efforts, and maintained by unequal and racialized legal, political, and economic systems which persist today (Robinson, 1983). We can therefore understand modern inequalities as embedded in consecutive iterations from the First Industrial Revolution—the 4IR still relies on the exploitation of labour through ghost workers and e-waste workers, extraction of knowledge though digital extractivism, and natural resources through technology companies' reliance on rare earth minerals (Mosco & Wasko, 1988; Agrawal et al., 2019; Keskin & Kiggins, 2021; Arrighi, 2008).

Conclusion

By outlining key differences between Global South and Global North AI discourses, we can identify gaps in current AI governance approaches for mitigating risks of AI systems, in ways that are relevant to a global population majority. Gaps discussed in this chapter include digital sovereignty as relevant to low- and middle-income countries, infrastructural and regulatory monopolies, harms associated with the labour and material supply chains of AI technologies, and commercial exploitation. Concomitantly, AI governance is subject to a historic amnesia that invisibilizes coloniality and eurocentrism embedded in AI infrastructures and governance processes. This further limits the understanding of present AI harms, and anticipatory capabilities within governance processes.

These gaps warrant a systemic restructuring of AI governance processes beyond current frameworks of inclusive AI governance. If AI risks are to be meaningfully mitigated at the societal and geopolitical levels, it is clear that inclusion is the first step of a long effort towards institutional reform that allows for adequate distribution of agenda-setting, decision-making, and resource power, as well as accountability structures for Global South and civil society actors.

Proposed steps for governance processes advanced by this chapter are as follows: (1) engage in a historical-geopolitical power analysis of structural inequality in AI governance and international legal frameworks; (2) identify mechanisms and protocols that mitigate "paradoxes of participation" and redress institutional barriers, in order to meaningfully engage with underrepresented stakeholder groups; and (3) co-construct and formalize roles for Global South actors to substantively engage in AI governance processes. These roles include the ability to act as a challenging function to exclusionary governance mechanisms, act as examples of alternative governance mechanisms, provide legitimate expertise in the interpretation of AI harms, and generate a political economy of resistance. Moving forward, the ability to materialize the aspiration of beneficial AI, mitigate AI harms, and develop appropriate guardrails, necessitates institutional reform, co-governance with Global South actors, South–South solidarity, and ways for people and communities undermined by institutional and commercial practices to design and use AI technologies that center their day-to-day realities and volition.

References

Abdala, María Belén, Ortega, Andrés, & Pomares, Julia. (2020). Managing the transition to a multi-stakeholder artificial intelligence governance. *G20 Insights*. https://www.g20-insights.org/policy_briefs/managing-the-transition-to-a-multi-stakeholder-artificial-intelligence-governance/.

Abraham, David. (2017). *The elements of power: Gadgets, guns, and the struggle for a sustainable future in the rare metal age*. Reprint edition. Yale University Press.

Acharya, Amitav. (2018). *The end of American world order*. Polity.

Ackerman, John. (2004). Co-governance for accountability: Beyond "exit" and "voice." *World Development 32*(3), 447–463. https://doi.org/10.1016/j.worlddev.2003.06.015.

Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. (2020). Outcome document: First draft of the Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000373434.

Agrawal, Ajay, Gans, Joshua, and Goldfarb, Avi. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *National Bureau of Economic Research*, February. https://doi.org/10.3386/w25619.

Ahiakpor, James C. W. (1985). The success and failure of dependency theory: The experience of Ghana." *International Organization 39*(3), 535–552. https://doi.org/10.1017/s0020818300019172.

Ahmed, Sara. (2012). *On being included: Racism and diversity in institutional life*. Duke University Press.

AI Ethics Guidelines Global Inventory. (2021). AI ethics guidelines global inventory by AlgorithmWatch. AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/.

Albert, Mathias, & Werron, Tobias. (2020). *What in the world?: Understanding global social change*. Bristol University Press.

Ali, Syed Mustafa. (2016). A brief introduction to decolonial computing. *XRDS: Crossroads, the ACM Magazine for Students 22*(4), 16–21. http://oro.open.ac.uk/46718/.

Anne Showstack Sassoon. (2014). *Gramsci and contemporary politics: Beyond pessimism of the intellect*. Routledge.

Anghie, A., & Chimni, B. S. (2003). Third world approaches to international law and individual responsibility in internal conflicts*. *Chinese Journal of International Law 2*(1), 77–103. https://doi.org/10.1093/oxfordjournals.cjilaw.a000480.

Arora, Payal. (2018). Decolonizing privacy studies. *Television & New Media 20*(4), 366–378. https://doi.org/10.1177/1527476418806092.

Arrighi, Giovanni. (2008). Historical perspectives on states, markets and capitalism, East and West. *The Asia-Pacific Journal 6*(1).

Bauder, Harald, & Mueller, Rebecca. (2021). Westphalian vs. indigenous sovereignty: Challenging colonial territorial governance. *Geopolitics*. 10.1080/14650045.2021.1920577.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). On the dangers of stochastic parrots: Can language models be too big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March, 610–623. https://doi.org/https://doi.org/10.1145/3442188.3445922.

Benjamin, Ruha. (2019). *Race after technology: Abolitionist tools for the new jim code*. Polity.

Berditchevskaia, Aleks, Malliaraki, Eirini, & Peach, Kathy. (2021). Participatory AI for humanitarian innovation: A briefing paper. *Nesta*. https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/.

Bhambra, Gurminder K., Gebrial, Dalia, & Nişancıoğlu, Kerem. (2018). *Decolonising the university*. Pluto Press.

Bird, Eleanor, Fox-Skelly, Jasmin, Jenner, Nicola, Larbey, Ruth, Weitkamp, Emma, & Weitkamp, Alan. (2020). The ethics of artificial intelligence: Issues and initiatives. European Parliament Scientific Foresight Unit. https://www.europarl.europa.eu/RegData/Etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf.

Birhane, Abeba. (2020). Algorithmic colonization of Africa. *SCRIPT-Ed 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389.

Bliss, Frank, & Neumann, Stefan. (2008). Participation in international development discourse and practice: "State of the art" and challenges. *INEF-Report*, 94. https://duepublico2.uni-due.de/receive/duepublico_mods_00027005.

Bottino, Celina. (2021). AI governance and policy in the Global South. Presented at the Virtual IEAI Speaker Series, March 25. https://ieai.mcts.tum.de/event/virtual-ieai-speaker-series-ai-governance-and-policy-in-the-global-south-with-celina-bottino/.

Brevini, Benedetta. (2022). *Is AI good for the planet?* Polity Press.

Browne, Simone. (2021). Acrylic, metal, blue and a means of preparation: Imagining and living black life beyond the surveillance state." *Microsoft.com*, October 27. https://www.microsoft.com/en-us/research/video/acrylic-metal-blue-and-a-means-of-preparation-imagining-and-living-black-life-beyond-the-surveillance-state/.

Buecker, Regina. (2003). Karl Marx's conception of international relations. *Glendon Papers*.

Bughin, Jacques, Seong, Jeongmin, Manyika, James, Chui, Michael, & Joshi, Raoul. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy." McKinsey Global Institute.

Campbell-Stephens, Rosemary. (2020) Global majority: Decolonising the language and reframing the conversation about race. *Global Majority 2020*. https://www.leedsbeckett.ac.uk/-/media/files/schools/school-of-education/final-leeds-beckett-1102-global-majority.pdf.

Capurro, Rafael. (2008). Information Ethics for and From Africa. *Journal of the American Society for Information Science and Technology* 59 (7): 1–9.

Capurro, Rafael. (2009). Intercultural information ethics: Foundations and applications. *Signo Y Pensamiento 28*(55), 66–79. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-48232009000200004.

Chan, Alan, Okolo, Chinasa T., Terner, Zachary, & Wang, Angelina. (2021). The limits of global inclusion in AI development. *ArXiv:2102.01265 [Cs]*, February. https://arxiv.org/abs/2102.01265.

Charnovitz, Steve. (2006). Accountability of non-governmental organizations in global governance." *NGO accountability politics, principles and innovations*. Routledge. https://doi.org/10.4324/9781849772099.

Chatham House. (2021). Reflections on building more inclusive global governance. April. https://www.chathamhouse.org/sites/default/files/2021-04/2021-04-15-reflections-building-inclusive-global-governance.pdf.

Cheikh, K. (2010). The Relationship between IP, Technology Transfer, and Development. *Intellectual Property Watch*, August 30. https://www.ip-watch.org/2010/08/30/the-relationship-between-ip-technology-transfer-and-development/

Chimni, B. S. (2006). Third World approaches to international law: A manifesto. *International Community Law Review 8*(1), 3–27. https://doi.org/10.1163/187197306779173220.

Cisneros, Nora Alba. (2018). "To my relations": Writing and refusal toward an indigenous epistolary methodology. *International Journal of Qualitative Studies in Education 31*(3), 188–196. https://doi.org/10.1080/09518398.2017.1401147.

Cleaver, Frances. (1999). Paradoxes of participation: Questioning participatory approaches to development. *Journal of International Development 11*, 597–612. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.461.2819.

Clutton-Brock, Peter, Rolnick, David, Donti, Priya L., & Kaack, Lynn H. (2021). Climate change and AI. *Global Partnership on AI Report*.

Coate, Roger A., Griffin, Jeffrey A., & Elliott-Gower, Steven. (2015). Interdependence in international organization and global governance. *Oxford Research Encyclopedia of International Studies*, December. https://doi.org/10.1093/acrefore/9780190846626.013.110.

Colloredo-Mansfeld, Rudi. (2002). Autonomy and interdependence in native movements: Towards a pragmatic politics in the Ecuadorian Andes. *Identities 9*(2), 173–195. https://doi.org/10.1080/10702890212205.

Comaroff, Jean, & Comaroff, John L. (2012). *Theory from the South, or, how Euro-America is evolving toward Africa*. Routledge, Taylor & Francis Group.

Connell, Raewyn. (2007). The Northern theory of globalization. *Sociological Theory 25*(4), 368–385. https://doi.org/10.1111/j.1467-9558.2007.00314.x.

Costanza-Chock, Sasha. (2020). *Design justice*. The MIT Press.

Couldry, Nick, & Mejias, Ulises Ali. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Crawford, Kate. (2021). *Atlas of AI*. Yale University Press.

Rivera Cusicanqui, S. (2012). Ch'ixinakax utxiwa: A reflection on the practices and discourses of decolonization. *South Atlantic Quarterly 111*(1), 95–109.https://en.wiktionary.org/wiki/gatopardismo.

D'Ignazio, Catherine, & Klein, Lauren F. (2020). *Data feminism*. MIT Press.

de Vicente, José Luis. (2013). The weight of the cloud. *CCCBLAB*, December 27. https://lab.cccb.org/en/the-weight-of-the-cloud/.

Dey, Bidit, & Ali, Faizan. (2016). A critical review of the ICT for development research. *ICTs in Developing Countries*, 3–23. https://doi.org/10.1057/9781137469502_1.

Disalvo, Carl. (2015). *Adversarial design*. MIT Press.

Dryzek, John S. (2012). Global civil society: The progress of post-Westphalian politics. *Annual Review of Political Science 15*(1), 101–119. https://doi.org/10.1146/annurev-polisci-042010-164946.

Edgerton, David. (2007). *The shock of the old: Technology and global history since 1900*. Oxford University Press.

Édouard Glissant, & Dash, J. Michael. (1999). *Caribbean discourse: Selected essays*. University of Virginia Press.

Elmi, Nima. (2020). Is big tech setting Africa back?" *Foreign Policy*, November 11. https://foreignpolicy.com/2020/11/11/is-big-tech-setting-africa-back/.

Escobar, Arturo. (2004). Beyond the Third World: Imperial globality, global coloniality and anti-globalisation social movements. *Third World Quarterly 25*(1), 207–230. https://www.jstor.org/stable/3993785.

Eubanks, Virginia. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fernández-Villaverde, Jesús, Mandelman, Federico, Yu, Yang, & Zanetti, Francesco. (2021). The "Matthew Effect" and Market Concentration: Search Complementarities and Monopsony Power. *Papers.ssrn.com*. February 18. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787787.

Fraser, Nancy. (2005). *Reframing justice*. Uitgeverij Van Gorcum.

Freire, Paulo. (1970). *Pedagogy of the oppressed*. Bloomsbury Academic.

Frey, William R., Patton, Desmond U., Gaskell, Michael B., & McGregor, Kyle A. (2018). Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured Twitter data. *Social Science Computer Review 38*(1), 42–56. https://doi.org/10.1177/0894439318788314.

Friedman, Batya, Kahn, Peter H., Borning, Alan, & Huldtgren, Alina. (2013). Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). https://doi.org/10.1007/978-94-007-7844-3_4.

Fuchs, Christian. (2009). Information and communication technologies and society. *European Journal of Communication 24*(1), 69–87. https://doi.org/10.1177/0267323108098947.

Gabriel, Iason. (2020). Artificial intelligence, values, and alignment. *Minds and Machines 30*(October), 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Gangadharan, Seeta Peña. (2013). Toward a deliberative standard: Rethinking participation in policymaking. *Communication, Culture & Critique 6*(1), 1–19. http://eprints.lse.ac.uk/67052/.

Garcia, Eugenio V. (2021). The international governance of AI: Where is the Global South? *The Good AI*, January 28. https://thegoodai.co/2021/01/28/the-international-governance-of-ai-where-is-the-global-south/.

Garfinkel, Ben. (2018). How sure are we about this AI stuff?" Presented at the Effective Altruism Global 2018, October 27. https://www.effectivealtruism.org/articles/ea-global-2018-how-sure-are-we-about-this-ai-stuff/.

GISWatch. (2020). *Global information society watch 2020 technology, the environment and a sustainable world: responses from the global south*. Alan Finlay (Ed.). APC. https://giswatch.org/sites/default/files/giswatch_2020_english_0.pdf.

Gopalakrishnan, Shanthi, & Santoro, Michael D. (2004). Distinguishing between knowledge transfer and technology transfer activities: The role of key organizational factors. *Papers.ssrn.com*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495508.

Graham, Mark. (2014). A critical perspective on the potential of the internet at the margins of the global economy. In *Society and the internet: How networks of information and communication are changing our lives* (pp. 201–318). Oxford University Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2448223.

Graham, Mark, De Sabbata, Stefano, & Zook, Matthew A. (2015). Towards a study of information geographies: (Im)Mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment 2*(1), 88–105. https://doi.org/10.1002/geo2.8.

Gramsci, Antonio. (1971). *Selections from the prison notebooks of Antonio Gramsci*. International Publishers (Original work 1891–1937).

Gray, Mary, & Suri, Siddharth. (2019). *Ghost work: How Amazon, Google, and Uber are creating a new global underclass*. Houghton Mifflin Harcourt.

Green, Ben. (2019). "Good" isn't good enough. In Proceedings of the AI for Social Good Workshop, Vancouver, Canada.

Greenpeace. (2020). Oil in the cloud: How tech companies are helping big oil profit from climate destruction. https://www.greenpeace.org/usa/reports/oil-in-the-cloud/.

Grosfoguel, Ramón. (2016). From "economic extractivism" to "epistemical extractivism" and "ontological extractivism": A destructive way to know, be and behave in the world. *Tabula Rasa 24*(24), 123–143. http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S1794-24892016000100006.

Gürses, Seda, Kundnani, Arun, & Van Hoboken, Joris. (2016). Crypto and empire: The contradictions of counter-surveillance advocacy. *Media, Culture & Society 38*(4), 576–590. https://doi.org/10.1177/0163443716643006.

Hattori, T. (2003). The moral politics of foreign aid. *Review of International Studies 29*(2), 229–247

Held, David, & Roger, Charles. (2013). *Global governance at risk*. Polity.

Helleiner, Eric. (2014, July–Sept.). Introduction. *Global Governance 20*(3), 359–360.

Hobson, J. M. (2012). The Eurocentric Conception of World Politics: Western International Theory, 1760–2010. In *Cambridge University Press*. Cambridge: Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/eurocentric-conception-of-world-politics/A2ECFA177E0199DADDCDBBD0295DF33D

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Classical ethics in A/IS. https://standards.ieee.org/Wp-Content/Uploads/Import/Documents/Other/Ead_classical_ethics_ais_v2.Pdf.

ILO. (2019). 50 million tonnes of potentially job creating e-waste discarded annually. *www.ilo.org*, January 25. https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_665137/lang--en/index.htm.

Irani, Lilly, & Silberman, Six. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *CHI 2013: Changing Perspectives*, Paris, France. http://crowdsourcing-class.org/readings/downloads/ethics/turkopticon.pdf.

Irani, Lilly, Vertesi, Janet, Dourish, Paul, Philip, Kavita, & Grinter, Rebecca E. (2010). Postcolonial computing. *Proceedings of the 28th International Conference on Human Factors in Computing Systems — CHI '10*. https://doi.org/10.1145/1753326.1753522.

Iyer, Neema, Achieng, Garnett, Borokini, Favour, & Ludger, Uri. (2021). Automated imperialism, expansionist dreams: Exploring digital extractivism in Africa. Pollicy. https://archive.pollicy.org/wp-content/uploads/2021/06/Automated-Imperialism-Expansionist-Dreams-Exploring-Digital-Extractivism-in-Africa.pdf.

Jaeger, Hans-Martin. (2007). "Global civil society" and the political depoliticization of global governance. *International Political Sociology 1*(3), 257–277. https://doi.org/10.1111/j.1749-5687.2007.00017.x.

Jameson, Shaz. (2019). Extractive logics as a systemic pattern. *Global Data Justice* (blog). April 13. https://globaldatajustice.org/2019-04-13-extractive-logics/.

Jasanoff, Sheila, & Hurlbut, J. Benjamin. (2018). A global observatory for gene editing. *Nature 555*(7697), 435–437. https://doi.org/10.1038/d41586-018-03270-w.

Jobin, Anna, Ienca, Marcello, & Vayena, Effy. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kak, Amba. (2020). The Global South is everywhere, but also always somewhere. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, February. https://doi.org/10.1145/3375627.3375859.

Kalluri, Pratyusha. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature 583*(7815), 169–69. https://doi.org/10.1038/d41586-020-02003-2.

Kalm, Sara, Strömbom, Lisa, & Uhlin, Anders. (2019). Civil society democratising global governance? Potentials and limitations of "counter-democracy." *Global Society 33*(4), 499–519. https://doi.org/10.1080/13600826.2019.1640189.

Karatasli, Sahan Savas. (2020). The politics of operations: Excavating contemporary capitalism. *Contemporary Sociology: A Journal of Reviews 49*(6), 527–529. https://doi.org/10.1177/0094306120963121r.

Keohane, Robert Owen, & Nye, Joseph S. (1989). *Power and interdependence*. Harpercollins.
Keskin, Tugrul, & Kiggins, Ryan David. (2021). *Towards an international political economy of artificial intelligence*. Palgrave Macmillan.
Khakurel, Jayden, Penzenstadler, Birgit, Porras, Jari, Knutas, Antti, & Zhang, Wenlu. (2018). The rise of artificial intelligence under the lens of sustainability. *Technologies 6*(4), 100. https://doi.org/10.3390/technologies6040100.
Kleven, Anthony. (2019). Belt and Road: Colonialism with Chinese characteristics. *www.lowyinstitute.org*, May 6. https://www.lowyinstitute.org/the-interpreter/belt-and-road-colonialism-chinese-characteristics.
Kovacs, A., & Ranganathan, N. (2019). Data sovereignty, of whom? Limits and suitability of sovereignty frameworks for data in India." Data Governance Network Working Paper 03.
van de Poel, I. R., & Kroes, P. A. (2014). Can technology embody values? *The Moral Status of Technical Artefacts* (pp. 103–124). https://research.tudelft.nl/en/publications/can-technology-embody-values.
Kulynych, B., Madras, D., Milli, S., Raji, I. D., Zhou, A., & Zemel, R. (2020). Participatory approaches to machine learning. In International Conference on Machine Learning Workshop.
Kurzweil, Ray. (2005). *The singularity is near: When humans transcend biology*. Duckworth.
Kwet, Michael. (2019). Digital colonialism: US empire and the new imperialism in the global south. *Race & Class 60*(4), 3–26. https://doi.org/10.1177/0306396818823172.
Legassick, Martin. (1974). South Africa: Capital accumulation and violence. *Economy and Society 3*(3), 253–291. https://doi.org/10.1080/03085147400000014.
Lindroth, Marjo, & Sinevaara-Niskanen, Heidi. (2018). *Global politics and its violent care for indigeneity: Sequels to colonialism*. Palgrave Macmillan.
Mahler, Anne Garland. (2017). Global South. *Oxford Bibliographies Online Datasets*, October. https://doi.org/10.1093/obo/9780190221911-0055.
Mahmoud, Mustafa. (2019). Stopping the digital ID register in Kenya—A stand against discrimination. *Namati* (blog), April 25. https://namati.org/news-stories/stopping-the-digital-id-register-in-kenya-a-stand-against-discrimination/.
Maldonado-Torres, Nelson. (2007). On the coloniality of being: Contributions to the development of a concept. *Cultural Studies 21*(2–3), 240–270.
Mantelero, Alessandro. (2018). AI and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review 34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017.
Marchetti, Raffaele. (2016). Global civil society. *E-International Relations*, December 28. https://www.e-ir.info/2016/12/28/global-civil-society/.
Mazzucato, Mariana. (2018). *The entrepreneurial state: Debunking public vs. private sector myths*. Public Affairs.
Mbembe, Achille, & Nuttall, Sarah. (2004). Writing the world from an African metropolis. *Public Culture 16*(3), 347–372. https://doi.org/10.1215/08992363-16-3-347.
McDowell, Caesar. (2016). Design for the margins. Presented at the TEDx Indiana University, June 9. https://interactioninstitute.org/design-for-the-margins/.
Mcglinchey, Stephen, Walters, Rosie, & Scheinpflug, Christian. (2017). *International relations theory*. E-International Relations Publishing. https://www.e-ir.info/publication/international-relations-theory/.
Mejias, Ulises Ali. (2020). To fight data colonialism, we need a non-aligned tech movement. www.aljazeera.com. September 8. https://www.aljazeera.com/opinions/2020/9/8/to-fight-data-colonialism-we-need-a-non-aligned-tech-movement.

Menski, Werner. (2006). *Comparative law in a global context: The legal systems of Asia and Africa*. Second Edition. Cambridge University Press.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel* April 8. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

Mhlambi, Sabelo. (2020). From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center Discussion Paper Series*, no. 2020-009. https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial.

Mignolo, Walter. (2012). *Local histories/global designs: Coloniality, subaltern knowledges, and border thinking*. Princeton University Press.

Milan, S. (2013). *Social Movements and Their Technologies*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137313546

Milan, Stefania, & Gutiérrez, Miren. (2015). Citizens' media meets big data: The emergence of data activism. *Mediaciones 11*(14), 120–133.

Milan, Stefania, & Treré, Emiliano. (2019). Big data from the South(S): Beyond data universalism. *Television & New Media 20*(4), 319–335. https://doi.org/10.1177/1527476419837739.

Milan, Stefania, & van der Velden, Lonneke. (2016). The alternative epistemologies of data activism. *Digital Culture & Society 2*(2). https://doi.org/10.14361/dcs-2016-0205.

Mohamed, Shakir, Png, Marie-Therese, & Isaac, William. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology 33* (July). https://doi.org/10.1007/s13347-020-00405-8.

Mosco, Vincent, & Wasko, Janet. (1988). *The political economy of information*. The University of Wisconsin Press.

Müller, Martin. (2018). In search of the Global East: Thinking between North and South. *Geopolitics 25*(3), 1–22. https://doi.org/10.1080/14650045.2018.1477757.

Murphy, James T., Carmody, Pádraig, & Surborg, Björn. (2014). Industrial transformation or business as usual? Information and communication technologies and africa's place in the global information economy. *Review of African Political Economy 41*(140), 264–283. https://doi.org/10.1080/03056244.2013.873024.

Mytton, David. (2021). Data centre water consumption. *NPJ Clean Water 4*(1), 1–6. https://www.nature.com/articles/s41545-021-00101-w.

Nash, Andrew. (2003). Third Worldism. *African Sociological Review/Revue Africaine de Sociologie 7*(1). https://doi.org/10.4314/asr.v7i1.23132.

Nayebare, Micheal. (2019). Artificial intelligence policies in Africa over the next five years. *XRDS: Crossroads, the ACM Magazine for Students 26*(2), 50–54. https://doi.org/10.1145/3368075.

Nedzhvetskaya, Nataliya, & Tan, J. S. (2021). The role of workers in AI ethics and governance. In J. Bullock (Ed.), *The Oxford Handbook on AI Governance*. Oxford University Press.

Neupane, Sujaya, & Smith, Matthew L. (2017). *Artificial intelligence and human development*. International Development Research Centre. https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949.

Nissenbaum, H. (2001). How computer systems embody values. *Computer 34*(3), 120–119. https://doi.org/10.1109/2.910905.

Nobrega, Camilia, & Varon, Joana. (2021). Big tech, gatopardismo and data colonialism. October 29. https://thefirethisti.me/2021/10/29/92-big-tech-gatopardismo-and-data-colonialism-with-camila-nobrega-and-joana-varon/.

Ochigame, Rodrigo. (2019). How big tech manipulates academia to avoid regulation. *The Intercept*, December 20. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/.

Ohajinwa, Chimere, Van Bodegom, Peter, Vijver, Martina, & Peijnenburg, Willie. (2017). Health risks awareness of electronic waste workers in the informal sector in Nigeria. *International Journal of Environmental Research and Public Health 14*(8), 911. https://doi.org/10.3390/ijerph14080911.

Ó hÉigeartaigh, Seán S., Whittlestone, Jess, Liu, Yang, Zeng, Yi, & Liu, Zhe. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology 33*(4), 571–593. https://doi.org/10.1007/s13347-020-00402-x.

Papin-Ramcharan, Jennifer. (2007). Open source software: A developing country view. In K. St. Amant and B. Still (Eds.), *Handbook of research on open source software: Technological, economic, and social perspectives* (pp. 93–101). IGI Global.

Parikka, Jussi. (2015). *A Geology of Media*. University of Minnesota Press.

Pasquinelli, Matteo, & Joler, Vladan. (2020). The nooscope manifested: AI as instrument of knowledge extractivism. *AI & SOCIETY 36*(November). https://doi.org/10.1007/s00146-020-01097-6.

Pathways for Prosperity Commission. (2019). Digital diplomacy: Technology governance for developing countries. Pathways for Prosperity Commission.

Png, Marie-Therese. (Forthcoming). *Paradoxes of participation in inclusive AI governance: Four key approaches from global south & civil society discourse*. UNESCO.

Piketty, Thomas. (2014). *Capital in the twenty-first century*. Arthur Goldhammer (Trans). The Belknap Press of Harvard University Press.

Pouliot, Vincent, & Thérien, Jean-Philippe. (2017). Global governance in practice. *Global Policy 9*(2), 163–172. https://doi.org/10.1111/1758-5899.12529.

Quijano, Aníbal. (2000). Coloniality of power and eurocentrism in Latin America. *International Sociology 15*(2), 215–32. https://doi.org/10.1177/0268580900015002005.

Quinn, John, Frias-Martinez, Vanessa, & Subramanian, Lakshminarayan. (2014). Computational sustainability and artificial intelligence in the developing world. *AI Magazine 35*(3), 36. https://doi.org/10.1609/aimag.v35i3.2529.

Rainie, Stephanie Carroll, Kukutai, Tahu, Walter, Maggie, Figueroa-Rodríguez, Oscar Luis, Walker, Jennifer, & Axelsson, Per. (2019). *Indigenous data sovereignty*. African Minds and the International Development Research Centre (IDRC). https://researchcommons.waikato.ac.nz/handle/10289/12918.

Raval, Noopur, Kak, Amba, & Calcaño, Alejandro. (2021). A new AI lexicon: Responses and challenges to the critical AI discourse. *Medium.com*, January 19. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-responses-and-challenges-to-the-critical-ai-discourse-f2275989fa62.

Rayment, P. B. W. (1983). Intra-"industry" specialisation and the foreign trade of industrial countries. *Controlling Industrial Economies*, 1–28. https://doi.org/10.1007/978-1-349-06340-6_1.

Ricaurte, Paola. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media 20*(4), 350–365. https://doi.org/10.1177/1527476419831640.

Ricaurte Quijano, Paola. (2020). Technology for life: Resistance from indigenous and urban communities in Mexico. *Technology, the Environment and a Sustainable World: Responses from the Global South, by GISWatch*. https://giswatch.org/sites/default/files/giswatch_2020_english_0.pdf.

Richardson, Helen, Tapia, Andrea, & Kvasny, Lynette. (2006). Introduction: Applying critical theory to the study of ICT. *Social Science Computer Review 24*(3), 267–273. https://doi.org/10.1177/0894439306287971.

Risse, Mathias. (2018). Human rights and artificial intelligence: An urgently needed agenda. *Revista Publicum 4*(1). https://doi.org/10.12957/publicum.2018.35098.

Robinson, Cedric J. (1983). *Black Marxism: The making of the Black radical tradition*. University Of North Carolina Press.

Romanoff, Mila. (2019). Building ethical AI approaches in the African context. *UN Global Pulse*, August 28. https://www.unglobalpulse.org/2019/08/ethical-ai-approaches-in-the-african-context/.

Rosa, Fernanda R., & Hauge, Janice Alane. (2020). GAFA's information infrastructure distribution: Implications for the Global South. *Papers.ssrn.com*, December 15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3749732.

Rosanvallon, Pierre, & Goldhammer, Arthur. (2008). *Counter-democracy: Politics in an age of distrust*. Cambridge University Press.

Saltelli, Andrea, Dankel, Dorothy J., Di Fiore, Monica, Holland, Nina, & Pigeon, Martin. (2022). Science, the endless frontier of regulatory capture. *Futures 135*(January), 102860. https://doi.org/10.1016/j.futures.2021.102860.

Sampath, P. (2021). Governing Artificial Intelligence in an Age of Inequality. *Global Policy* 12(S6):21–31. https://doi.org/10.1111/1758-5899.12940

Sassen, Saskia. (2017). Predatory formations dressed in Wall Street suits and algorithmic math. *Science, Technology & Society 22*(1), 1–15. https://doi.org/10.1177/0971721816682783.

Satariano, Adam. (2021). Europe proposes strict rules for artificial intelligence. *The New York Times*, April 21. https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html.

Schiff, Daniel, Borenstein, Jason, Biddle, Justin, & Laas, Kelly. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society 2*(1), 31–42. https://doi.org/10.1109/tts.2021.3052127.

Scholte, Jan Aart. (2004). Civil society and democratically accountable global governance. *Government and Opposition 39*(2), 211–233. https://doi.org/10.1111/j.1477-7053.2004.00121.x.

Sengers, Phoebe, Boehner, Kirsten, David, Shay, & Kaye, Joseph 'Jofish'. (2005). Reflective design. *Proceedings of the 4th Decennial Conference on Critical Computing between Sense and Sensibility — CC '05*. https://doi.org/10.1145/1094562.1094569.

Sheng, Li, & do Nascimento, Dmitri Felix. (2021). The Belt and Road Initiative (BRI): The South–South Cooperation (SSC) with Chinese characteristics. *The Belt and Road Initiative in South–South Cooperation* (pp. 1–43). https://doi.org/10.1007/978-981-16-6357-4_1.

Singh, Ranjit. (2021). Mapping AI in the Global South. *Medium*, January 26. https://points.datasociety.net/ai-in-the-global-south-sites-and-vocabularies-e3b67d631508.

Singh, Ranjit, & Guzmán, Rigoberto Lara. (2021). Parables of AI in/from the Global South. *Data & Society*, September 7. https://datasociety.net/announcements/2021/07/13/call-for-participants-parables-of-ai-in-from-the-global-south/.

Smith, Adrian, & Stirling, Andy. (2008). Social-ecological resilience and socio-technical transitions: Critical issues for sustainability governance. STEPS Centre. https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/2438/Social-%20ecological%20resilience%85..pdf?sequence=1.

Smogorzewski, Kazimierz. (1938). Poland's foreign relations. *The Slavonic and East European Review 16*(48), 558–571.

Smuha, Nathalie A. (2020). Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea. *Philosophy & Technology 34*(May), 91–104. https://doi.org/10.1007/s13347-020-00403-w.

South Centre. (2020). Submission by the South Centre to the draft issues paper on intellectual property policy and artificial intelligence (WIPO/IP/AI/2/GE(20/1). https://www.southcentre.int/wp-content/uploads/2020/02/Submission-by-SC-to-the-Draft-Issues-Paper-on-IP-Policy-and-AI.pdf.

Stallman, Richard. (2002). On hacking. http://stallman.org/articles/on-hacking.html.

Star, Susan Leigh. (1990). Power, technology and the phenomenology of conventions: On being allergic to onions. *The Sociological Review 38*(1 suppl), 26–56. https://doi.org/10.1111/j.1467-954x.1990.tb03347.x.

Star, Susan Leigh, & Strauss, Anselm. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW) 8*(1–2), 9–30. https://doi.org/10.1023/a:1008651105359.

Strubell, Emma, Ganesh, Ananya, & Mccallum, Andrew. (2019). Energy and policy considerations for deep learning in NLP. https://arxiv.org/pdf/1906.02243.pdf.

Tapia, Danae, & Peña, Paz. (2020). White gold, digital destruction: Research and awareness on the human rights implications of the extraction of lithium perpetrated by the tech industry in Latin American ecosystems. In GISWatch (Ed.), *Technology, the environment and a sustainable world: Responses from the Global South*. https://giswatch.org/sites/default/files/giswatch_2020_english_0.pdf.

Taylor, Keeanga-Yamahtta. (2019). *Race for profit: How banks and the real estate industry undermined Black homeownership*. University of North Carolina Press.

Taylor, Linnet. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society 4*(2), 205395171773633. https://doi.org/10.1177/2053951717736335.

Teevan, Chloe. (2021). Building strategic European digital cooperation with Africa. ECDPM. https://ecdpm.org/wp-content/uploads/Building-Strategic-European-Digital-Cooperation-With-Africa-ECDPM-Briefing-Note-134-2021.pdf.

Thatcher, Jim, O'Sullivan, David, & Mahmoudi, Dillon. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data." *Environment and Planning D: Society and Space 34*(6), 990–1006. https://doi.org/10.1177/0263775816633195.

Torres, Guillén. (2017). Taking a look at institutional resistance to citizen empowerment. *DATACTIVE* (blog), February 10. https://data-activism.net/2017/02/blog-taking-a-look-at-institutional-resistance-to-citizen-empowerment-through-data/.

Treré, Emiliano. (2018). From digital activism to algorithmic resistance. In *The Routledge companion to media and activism* (pp. 367–375). Routeledge.

Ulnicane, Inga, Eke, Damian Okaibedi, Knight, William, Ogoh, George, & Stahl, Bernd Carsten. (2021). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews 46*(1–2), 71–93. https://doi.org/10.1080/03080188.2020.1840220.

UNCTAD. (2013). Information Economy Report 2013: The cloud economy and developing countries. United Nations Conference on Trade and Development.

van Stam, Gertjan. (2020). Power inequities: Observations on the development of information and communication technologies, from an African place. *12th ACM Conference on Web Science Companion*, July. https://doi.org/10.1145/3394332.3402830.

Vawda, Yousuf. (2021). The TRIPS COVID-19 waiver, challenges for Africa and decolonizing intellectual property. *South Centre Policy Brief 99*(August). https://www.southcentre.int/policy-brief-99-august-2021/.

Velluet, Quentin. (2021). Can Africa salvage its digital sovereignty? *The Africa Report.com*, April 16. https://www.theafricareport.com/80606/can-africa-salvage-its-digital-sovereignty/.

Wade, Robert H. (2004). On the causes of increasing world poverty and inequality, or why the "Matthew Effect" prevails. *New Political Economy* 9(2), 163–188. https://doi.org/10.1080/1356346042000218050.

Wade, Robert H. (2013). Capitalism and democracy at cross-purposes. *Challenge* 56(6).

Walter, Maggie, Kukutai, Tahu, Carroll, Stephanie Russo, & Rodriguez-Lonebear, Desi. (2020). *Indigenous data sovereignty and policy*. Routledge.

Wardrip-Fruin, Noah, & Montfort, Nick. (2003). *The new media reader*. MIT Press.

Weiss, Thomas G. (2016). Rising powers, global governance, and the United Nations." *Turkey in Global Governance: Searching for Alternatives between the West and the Rest 1*(2), 7–19.

Weiss, Thomas G., & Abdenur, Adriana Erthal. (2014). Introduction: Emerging powers and the UN—What kind of development partnership? *Third World Quarterly* 35(10), 1749–1758. https://doi.org/10.1080/01436597.2014.971583.

Werner, David, & PROJIMO. (1998). *Nothing about us without us: Developing innovative technologies for, by and with disabled persons*. Healthwrights.

Whittaker, M. (2021). The Steep Cost of Capture. *ACM Interactions* (XXVIII), 50. Retrieved from https://interactions.acm.org/archive/view/november-december-2021/the-steep-cost-of-capture.

WHO. (2021). Children and digital dumpsites: Smaller hands, cheaper labour—the crisis of e-waste and children's health. *WHO*, June 15. https://www.who.int/news-room/questions-and-answers/item/children-and-digital-dumpsites-e-waste-and-health.

WIPO. (2004). *Proposal by Argentina and Brazil for the Establishment of a Development Agenda for WIPO WO/GA/31/11*. Geneva: WIPO. Retrieved from WIPO website: https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=31737.

Working Group on Trade Treaties and Access to Medicines. (2020). Open letter to TRIPS council members: Remove barriers to access for critical COVID-19 supplies. Global Development Policy Center, November 12. https://www.bu.edu/gdp/2020/11/18/open-letter-to-trips-council-members/.

Yuan, Li. (2018). How cheap labor drives China's A.I. ambitions. *The New York Times*, November 25. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html.

Zuboff, Shoshana. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.

CHAPTER 49

NATO'S ROLE IN RESPONSIBLE AI GOVERNANCE IN MILITARY AFFAIRS

ZOE STANLEY-LOCKMAN AND LENA TRABUCCO

Introduction

The introduction of artificial intelligence (AI) as a general-purpose technology has prompted analysts and researchers to reconsider the implications for warfare. As this *Handbook* edition illustrates, AI has, and will continue to, shape global dialogue, policy, and governance structures in international politics, including for future military operations.

In this chapter, we explore a role for the North Atlantic Treaty Organization (NATO) in the emerging military AI governance architecture. NATO (or the Alliance) is a military and political alliance among 30 contributing member states that are committed to collective security. Much of NATO's original purpose and current core tasks arguably leave the Alliance's role uncertain in international governance regimes contending with the impact of emerging technology on international politics.[1] As global powers compete for the economic and military capabilities that AI can offer, the Alliance has the enormously challenging task of navigating varying political realities and capabilities of Allies, all while effectively recalibrating strategic relationships in the coming years. Recognizing technological change as a key variable, NATO has begun to adapt its organizational composition and strategic footing to increase the Alliance's capacity to meet emerging security challenges for military capability development trends of both its own members and those of competitors or adversaries.

New power distributions around AI and adjacent dual-use technologies are among the motivating factors causing the Alliance to reconsider whether its technological superiority may be threatened in the years ahead, as reflected in the 2019 Emerging and Disruptive Technologies (EDTs) Roadmap[2] and, more recently, the NATO 2030 process.[3] NATO navigates these changes and then approaches AI-accelerated changes to the international

security environment in a highly political context. Notably, in 2019, French President Emmanuel Macron surprised many European counterparts by declaring NATO "brain-dead," a warning wrapped in an even larger warning of trans-Atlantic security divisions.[4] The critique that NATO is a "brain-dead" or "irrelevant" institution has existed in some form since the end of the Cold War.[5] As NATO combats global perceptions of organizational irrelevance, there is a reason to push for bureaucratic adaptation to better manage technology-driven changes in the future. As such, despite some warnings to the contrary, Allies have an incentive to keep NATO a relevant military institution and ensure that it adapts to emerging threats and for future military contexts. The comment from President Macron helped prompt the NATO 2030 agenda, which is currently taking shape to increase the Alliance's role as a political actor and as an organization with a greater focus on EDTs.[6]

As NATO bodies and Allies prepare for the impact of AI on future military operations, the Alliance has its own responsibility to steward AI in ways that, *inter alia*, promote cohesion between democratic countries, prevent risks, shore up interoperability, project deterrence, and ensure stability.[7] To achieve these aims, cooperation and alignment are critical for the Alliance to maintain a competitive edge and promote further innovation in alignment with shared values.

With these incentives, we argue that an examination of NATO as a governance stakeholder is due to complement other literature on how humans, social structures, and institutions impact how technology develops. More specifically, this chapter borrows from two fields of scholarship that set the theoretical foundations for how institutions such as NATO impact technological trajectories, and thus have a responsibility to govern the technology accordingly. The two fields—science, technology, and society (STS) studies and military innovation literature—have different parameters, but both explore key questions that help establish the ways in which institutions exert their influence on the development, deployment, and diffusion of technologies like AI. We argue that this influence is a form of institutional power, building on Seth Lazar's definition of governance in this handbook. Lazar writes that governance is "the use of power to make, implement, and enforce the constitutive norms of an institution."[8] In the context of this definition, this chapter examines AI governance as an instrument of power linking NATO's responsibility and capacity to shape the future security environment in parallel to its own organizational interests.

To be sure, NATO is far from the only institution that impacts military AI governance and its security implications. Indeed, international technology governance is inherently complex because it includes diverse stakeholders in a system of "organizations, regimes, and other forms of principles, norms, regulations, and decision-making procedures" with a shared interest and responsibility in a given issue-area of world politics.[9] Existing discussions of the impact of AI on international security have looked to nation-states, regional institutions like the European Union (EU), or international bodies like the United Nations Convention on Certain Conventional Weapons (UN CCW) for discussions on the military governance of AI.[10] Without expanding on the role of these other stakeholders, this chapter begins to explore pressing questions for NATO and international relations scholars that illustrate NATO's role in AI governance, which has not had a comprehensive analysis.[11]

To begin to fill this gap, the analysis in this chapter centers on two AI governance mechanisms that NATO has at its disposal, and subsequently explores the Alliance's capacity to use these mechanisms to exert its influence in key pillars of AI governance. Of the many possible AI governance mechanisms for NATO, this chapter offers a deeper

assessment of two: (1) strategic and policy planning and (2) standards and certification. We fashion these mechanisms as primary components that connect NATO technology governance measures and responsible AI use.[12] To illustrate NATO's capacity to govern AI, we then examine three pillars, or foundational issue areas, which we believe represent critical elements of technology governance. We argue that, within each pillar, NATO is uniquely situated to facilitate cooperation via its governance mechanisms, with a view to shaping the future of AI for the Alliance and maintaining a competitive edge. Each pillar—(1) ethics and values, (2) legal norms, and (3) security and safety—is an area where researchers and analysts have acknowledged significant governance challenges, both at a national level and for international organizations like NATO. Each pillar, discussed in depth below, illustrates NATO's potential as a governance stakeholder that can encourage multinational alignment on policy and standards for safer and better outcomes in future operations.

The rest of this chapter continues as follows. First, we establish how STS studies and scholarship on military innovation focus on different aspects of technological advancement and governance outlooks. Second, this theoretical basis is applied to NATO to provide readers with an understanding of the institution's entities and responsibilities related to AI governance. Next, the chapter discusses ways that NATO can leverage these mechanisms to ensure responsible use of AI in military operations based on ethics, law, safety, and security. Finally, the chapter concludes with reflections on NATO's AI governance tools and, more broadly, roles for international organizations in the AI governance space.

AI Governance and Military Affairs: Tensions in Existing Literature

Academic literature has long grappled with the intersection of emerging technology and security organizations.[13] Two branches of literature that tackle core questions of technological trajectories and its relationship to human and social structures—a critical question of governance for military technology—are STS studies and military innovation scholarship. Although the theoretical approaches in STS and military innovation studies differ, they both share the important assumption that technology does not have its own innate logic, and instead measure technological change by its impact on social structures and interactions with humans. In other words, both fields treat technology as an enabler in broader structures. The term technology is ubiquitous enough that it does not have a single definition, but it is often defined in relation to human intention and purpose. Alex Roland describes technology as a "purposeful human manipulation of the material world" to "serve some human purpose."[14] If extending this basic idea of technology to technological innovation, then both STS studies and military-innovation scholarship lend relevant criteria.

Both academic fields are also relevant because, in the policy space, AI governance stakeholders are pursuing responsible research and innovation (RRI), which comes from STS studies, and defense stakeholders are similarly focused on responsible innovation and responsible use. More traditionally, the direct study of military adoption of technology is considered in the separate scholarship of military innovation, which includes a school of thought that focus on cultural and organizational factors. Between these two fields, an

interdisciplinary approach is helpful here to carry STS approaches to AI governance, including RRI, over to the space of military innovation.

However, this is complicated by the reality that military organizations that see technological superiority as a core element of deterrence and defense, including NATO, engage in forms of technological determinism that STS scholars squarely reject. Respective views on technological determinism—which considers that technology shapes society as a largely autonomous process with limited human agency—thus creates a tension for governance prospects.[15] To spotlight the aspects of military innovation related to governance, this section briefly expands on the overlaps and tensions between STS and military innovation literature.

Science, technology, and society (STS) studies

STS studies is helpful to understand how technologies such as AI develop relative to the human, social, and political structures that shape it, rather than as an independent entity to which humans have to adapt.[16] In this vein, AI is not just a computational process involving software, hardware, and data,[17] so much it is a socio-technical system that encompasses "human, social, and organizational factors."[18] Together, these factors enable a focus on the trajectory of technological development relative to social structures and power dynamics. STS scholars have also helped develop RRI frameworks that seek to guide technological development in anticipatory, participatory, and adaptive frameworks to achieve desirable outcomes and prevent undesirable ones.[19] RRI is a structured approach to innovation in which stakeholders identify and act on their "collective commitment of care for the future through responsive stewardship of science and innovation in the present."[20] It drives civilian AI ecosystems for NATO Allies that will also indirectly affect NATO.[21]

Responsible stewardship, or governance, of science and technology (S&T) requires stakeholders to change their approaches to technological development as the circumstances themselves change.[22] In his book *The Social Control of Technology*, David Collingridge identified the double bind that makes technology governance (what he then referred to as social control) difficult: exerting social control or governing nascent technology is easy, but impossible because its evolution and eventual impacts are unknowable, and by the time the technology matures and its impact is realized, entrenched decisions will make future control more difficult.[23] For now, AI remains a relatively immature technology, meaning circumstances will change as knowledge emerges and norms progressively develop. Collingridge also suggested the necessity of "corrigibility of innovation," which refers to the "capacity to change shape or direction in response to stakeholder and public values and changing circumstances."[24] When applied to current RRI frameworks, the concept of corrigibility obligates governance stakeholders to shape the trajectory of a technology's development and impact in ways based on social structures, both in anticipation of change and in response to decisions made in error.[25] In short, stakeholders have to adopt corrigible practices to responsibly govern technology as it develops, and thus must claim their agency in guiding innovation even as technological development appears increasingly entrenched in previously made decisions and their subsequent outcomes.

This is important for AI governance because technological advancement is making AI-accelerated risks clearer, including in the military space. Risks—especially as related to

AI-enabled autonomous systems, poisoning of information environments, cyberattacks, unpredictable failure modes, and emergent behavior—will evolve in form and scale as the technology matures and diffuses. If AI evolution means more entrenchment and less corrigibility, the STS foundations remind governance stakeholders how to course-correct and adapt to changing risk assessments and the overall impact of AI in the international system.[26]

Nevertheless, while STS scholars study how decision-making that shapes the trajectory of technological innovation becomes entrenched, the field largely rejects the premise of technological determinism. Maintaining the centrality of human agency, as exerted also through social structures and institutions, is antithetical to determinist perspectives on technology developing on its own path independent of intervention. As Allan Dafoe, another contributor to this *Handbook*, has argued, the STS academic community's refusal to engage with technological determinism severely limits STS applicability to empirics.[27] As discussed below, this has implications for the ability of the STS field to impart responsibility to governance stakeholders in an area such as AI and international security.

Military innovation literature

The scholars that examine the way that military stakeholders manage technology and shape its development trajectory predominantly write on military innovation.[28] These scholars measure technology adoption in changes to doctrine, organizational structures, and operational concepts, rather than seeing the technology as an end in and of itself.[29] From this perspective, technology subsequently shapes human and social structures and organizations. To take an example similar to Roland's definition of technology, Jonathan Shimshoni's concept of "military entrepreneurship" involves the active manipulation of technology, doctrine, and war plans.[30] In this sense, new technology adoption has tangible and observable effects on the operational environment. Similarly, Thomas Mahnken illustrates that military services shape technology to their respective purposes, rather than the other way around.[31] The purpose that this manipulation, or molding, of technology serves is the creation, and ideally sustainment, of a comparative military advantage.[32]

Still, the way that this military advantage is defined is relevant here because metrics of success differ from other scholarship dealing with innovation. Military innovation importantly constitutes the relationship and social structures that form between technology and military bureaucracies. Yet as a field, it does not necessarily extend these relationships to their status as stakeholders in wider technology governance regimes.[33] For instance, in his review of the different schools of military innovation, Adam Grissom offers a consensus definition of military innovation that inherently links it to effectiveness in the battlespace.[34] Grissom clarifies that "measures that are administrative or bureaucratic in nature, such as acquisition reform, are not considered legitimate innovation unless a clear link can be drawn to operational praxis."[35] This reinforces the idea that technology on its own does not constitute an innovation if it is not observable in military operational practice or in battlefield advantage.

The relatively narrow operational focus of military innovation scholarship means that management structures miss out on some of the uses of military power implicit in the governance of military technology. This means that both the bureaucratic entrenchment of

technological advancement *and* the literature focusing on it do not necessarily address governance as an instrument of power in the military context. This may make sense for purely military technologies, but whether it discounts the agency that military bureaucracies have in governance of a pervasive, general-purpose technology like AI is worth separate consideration. As such, the operational measurement of the adoption and diffusion of technology as an instrument of military power likewise limits an understanding of how military technology management structures relate to governance.

Implications for military AI governance

Overall, STS offers much of the necessary groundwork for governance mechanisms and the impact of social structures on technology governance; however, it refuses to engage with technological determinism, or the independent influence of technology, that is often a driving force in military innovation. Recognizing that a comprehensive governance regime also needs to transpose to stakeholders that are engaged in the practices of governing AI, this study on NATO sees military innovation scholarship as a helpful complement to apply these STS foundations to practitioners' perspectives. But scholarship on military innovation also has its own flaws, in that it looks at the management of technology exclusively formulated to exploit a comparative operational advantage. Measuring military innovation in relation to operational praxis makes sense to detect how military adoption of technology impacts operational excellence and upstream impacts on military strategy, but also makes it challenging for the empirics to apply to non-operational ways that military organizations exert their influence. Non-operational influence includes governance, the core topic that this chapter addresses.

Despite differences, borrowing from the layered frameworks in STS and military innovation studies still helps contextualize innovation trajectories. Indeed, select scholars have attempted to bridge the gap between the social constructivist angle in STS studies and the technologically "optimistic"[36] assumptions that frame technologically deterministic undercurrents, as seen in case studies on military innovation.[37] Thomas Hughes examined these undercurrents in the defense sector as part of his theory of "technological momentum,"[38] which argued military organizations are subject to inaction in S&T decision-making because the entrenchment of previous investments and decisions constrain the course of future technological development. Steven Fino expands on Hughes with the idea that "technological dislocations," are an alternative reconciliation mechanism that acknowledges that technological determinism may operate beneath the surface of a technology's maturation trajectory, while still allowing for socially driven perturbations that "dislocate" the "otherwise logical evolutionary patterns" of that technology.[39]

Dafoe similarly attempts to widen the scope of technological determinism by placing it as an endpoint on a spectrum, with social constructivism on the other end. The purpose of this spectrum is to create the space for engagement with disciplines that heavily emphasize power dynamics, including military affairs and business in what he terms "military-economic adaptationism."[40] Unfortunately, both Fino and Dafoe concede that attributing agency and causality to technological developments are best "conducted after the fact"[41] or "on longer timescales," respectively.[42] AI governance cannot benefit from such hindsight, as it is fundamentally a question of how to project and adapt to forces of ongoing

change. For governance, this inertia places military organizations at odds with the responsiveness required to guide responsible technology governance frameworks. Our aim is not to reconcile these differences in this chapter, but rather to highlight how they frame one current governance challenge for military stakeholders such as NATO: how can they engage with the socio-technical foundations in RRI frameworks to shape, adapt to, and respond to technology-accelerated changes, while simultaneously pursuing their traditional aims of adopting technology to deter and defend?

On this note, it is worth mentioning that NATO itself has historically convened scholars from both STS and military innovation backgrounds to understand socio-technical changes to their operating environment.[43] The Alliance also takes socio-technical factors into account in its S&T work on emerging technologies—including human–systems integration, technology monitoring, and forecasting work.[44] This interest in socio-technical systems relating to effectiveness suggests scope for the Alliance to leverage technology governance as an instrument of its influence, as picked up in the next section.

NATO's Mechanisms to Govern AI

NATO's increasing interest in EDTs introduces the need to consider how governance priorities can help reinforce the Alliance's influence. The STS and military innovation literature provide the theoretical foundations for NATO's stewardship of AI as they place attention on "the role that institutions play in shaping technological trajectories."[45] As AI development continues, the actions that NATO and its members take will have important implications for their capacity to adopt, respond to, and shape their future operating environment. Particularly for democracies, this confers to military stakeholders a dual responsibility to prevent and manage risks, as well as to proactively shape their approach to technological development anchored in democratic values and security. As a multinational alliance with an incentive to drive cooperation and alignment, NATO is situated to define and operationalize norms, as well as promote standards that help shape the contours of future military effectiveness and technological competition.

In a RRI framework, not only is this an institutional role, but it also becomes an institutional responsibility. To apply this responsibility to NATO's stewardship of AI, the institutional interplay between technology, structure, and concepts is a form of socio-technical system with important implications for AI governance because they link the ways that an institution uses its power to adopt and shape AI trajectory to its respective ends.

Already, several mechanisms are built into military bureaucracies to ensure that technology is adopted in alignment with responsible engineering practices and responsible state behavior.[46] The Alliance is organized to harmonize between Allies so that their contributions enhance military effectiveness and political cohesion between like-minded democracies. We argue that these effectiveness-centric mechanisms likewise empower NATO to exert its influence in technology governance. More specifically, this entails the Alliance helping steward technological development for a more predictable strategic environment and enhanced democratic clout around the exploitation of technology reinforcing rule of law. For NATO, we focus on strategic and policy planning, as well as standards and certification because they reflect the Alliance's particular strengths and interests in S&T. These practices are relevant to

governance insofar as they exemplify an institution's power to shape the trajectory of technological development—but this selection is by no means exhaustive.[47]

Instead, our aim is to explore how these mechanisms are operationalized at the Alliance level. In this vein, Table 49.1 dissects the role that its various bodies play in managing technology, promulgating and operationalizing standards, and leading change through policy. The role of NATO in this equation is largely shaped by its members' own approaches to technology, and member-state-driven processes are complemented by "policy entrepreneurs" and technical experts in the International Staff and related bodies.[48] Table 49.1 does not list the ways that AI affects the various functions of NATO, but rather spotlights the entities that together operationalize AI governance through cumulative processes on policy and standardization.

Strategic and policy planning

NATO structures around strategic and policy planning both set Allied ambitions and priorities and have the competency to implement them through its many consultative bodies, coordination formats, and albeit to a lesser extent, technology foresight capacities. NATO has facilitative power among Allies, both for defense planning and for the conduct of operations. A cornerstone in modern architecture of international security is coalition warfare—or, more broadly, joint operations. Working with military partners has become a critical feature of modern security policy, where there is more power in enhancing numbers, but also in having allies that lend political and practical legitimacy to deterrence and operations.[49] NATO is vital to that effort for many reasons, but also because NATO's facilitative power is significant to promote coordination and cooperation. Simply put, partners and allies are a necessary feature of modern military behavior, and strategic and policy planning are necessary functions to encourage and underpin cohesion in alliance settings. This is important for AI governance because the nature of AI poses new strategic challenges and will require multilateral approaches and some degree of cohesion to effectively incorporate RRI frameworks in policy planning. As such, the necessity of working with security partners extends to the AI-policy frontier.

A number of NATO entities carry out strategic and policy planning, recognizing the importance of policy alignment to sustain political strength and military effectiveness. As relates to S&T, allies' representations to NATO, defense ministries, and policy entrepreneurs from the relevant entities summarized in Table 49.1 support and negotiate how the Alliance approaches EDTs. NATO's strategic documentation and forward-looking policy analysis incorporates hints of technological determinism, including noting how technological change *inevitably* shapes the future strategic and operating environment. Further, the connections between technology and competitive advantage over adversaries and competitors are embodied in the Alliance's desire to maintain its "technological edge" as the "foundation upon which NATO's ability to deter and defend against potential threats ultimately rests."[50] This places technology squarely within NATO's core purpose of deterrence and defense—and while this signals NATO's express commitment to technology through these channels, this reliance on technology also obscures whether NATO's governance capacity will be adaptive, anticipatory, or participatory. This position of technological determinism may result in more limitations for AI governance.

Table 49.1 NATO entities relevant to AI governance

NATO Entity	General Mission	Relevance to AI Governance
Allied Command Transformation (ACT)	One of two strategic commands (other being Allied Command Operations); leading adaptation and transformation efforts related to operational concepts, structures, and interoperability	Significant focus on ways AI impacts future military context from 2014 onward; developed Emerging and Disruptive Technologies Roadmap and led away day for North Atlantic Council to help create momentum for NATO AI agenda in 2018
Command, Control and Consultation Board (C3B)	Senior multinational policy body reporting to North Atlantic Council and Defence Planning Committee on information sharing and interoperability, including cyber defense, information assurance, and joint ISR	Coordination of AI-related policies, including TEVV frameworks and responsible development processes outside of NATO's remit
Communications and Information Agency (NCI Agency)	Agency focused on Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR) technology and communications capabilities for decision-making and mission support	Acquisition and experimentation of software and AI systems
Conference of National Armaments Directors (CNAD)	Senior committee of armaments directors who meet biannually to promote armaments cooperation and harmonize military requirements	Align acquisition policies and processes that focus on procurement and sustainment of AI systems (for instance, if procurement guidance includes legal and ethical reviews or stage-gating)
Defence Innovation Accelerator for the North Atlantic (DIANA)	Civil-military technology accelerator announced in June 2021 to be stood up by 2023 *[NB: NATO Innovation Fund also announced; governance of venture capital-styled fund not yet clear]*	Fund and coordinate activities on TEVV for emerging and disruptive technologies, including AI
Innovation Board	Board composed by senior staff and chaired by Deputy Secretary General to enable NATO staff to understand implications of new technology and innovation *[NB: not a decision-making committee]*	Disseminate NATO responsibilities in AI governance to stakeholders across the Alliance so they can understand new risks, implications, and opportunities
Innovation Unit and Data Policy Unit	Established in late 2019/early 2020 in the Emerging Security Challenges Division at NATO HQ to facilitate innovation ecosystem internally and externally	Implementation of forthcoming NATO AI Strategy and Emerging and Disruptive Technologies Roadmap, Data Exploitation Framework Policy, forthcoming AI Strategy

(*continued*)

Table 49.1 Continued

NATO Entity	General Mission	Relevance to AI Governance
North Atlantic Council (NAC)	Political decision-making body overseeing all political and military processes; Defence Policy and Planning Committee also responsible for defense planning on behalf of the NAC, including coordination of the NATO Defence Planning Process	Main forum for member states to address AI governance priorities, including setting ambition on ethical and legal basis for AI exploitation in military affairs and setting agenda for operationalization of principles, norms, and standards
Office of Legal Affairs (OLA)	Provide legal advice to Secretary General and International Staff on policy issues, legal defense of Alliance's interests, and ensure compliance for multinational operations	Allied compliance with relevant international legal regimes for AI (including international humanitarian law), coalition legal interoperability, national and international litigation
Science & Technology Board (STB) and Science and Technology Organization (STO)	STB: "promote NATO-wide coherence of NATO S&T" STO: entity with large network focused on maintaining "NATO's scientific and technological advantage by generating, sharing, and utilizing advanced scientific knowledge, technological developments and innovation to support the Alliance's core tasks," as per 2018 S&T Strategy	Complement ACT's work from 2014–2018 to provide technical baseline for impact of AI on future military operations and strategic context [Subset: Centre for Maritime Research and Experimentation—which is working on C2 systems for unmanned systems a basis for more automated or AI-enabled C2 developments—also under STO auspices]
NATO Standardization Office (NSO)	Independent office that leads standardization activities under the Committee for Standardization and support NATO Defence Planning Process	Promulgation of standards on practices related to documentation, safety, security, and ethics for training data and models, related to governance practices, etc.

Standards and certification

To maintain its relevance in a security architecture increasingly concerned with the way that technology shifts power dynamics and scales threats to international security, NATO has an incentive to foster cooperation, promote standards of practice, and incentivize Allied AI harmonization. It is strategically salient to facilitate a dialogue and engagement among Allies on AI, but it is practically important to use NATO's position to facilitate Allied cooperation regarding standards to project the Alliance's ability to interoperate in future operations. NATO standards aim to enhance interoperability among partners and successful implementation of strategy.

More specifically, standards and certification are used to establish and implement requirements aligned with safe development and responsible use of technology. In addition to purely technical standards, NATO has operational standards that specify "conceptual, organizational or methodological requirements to enable materiel, installations, organizations

or forces to fulfil their functions or missions."[51] In line with the definitions from STS and military innovation scholarship, standards can thus be seen as a mechanism to translate responsibility-derived state and organizational AI policy into actionable functions. In fact, NATO has set certain standards for the Allies and these standards subsequently become the norm.

Within NATO, it is the NATO Standardization Office (NSO) that coordinates thousands of experts to align technological development with military requirements that can help enhance effectiveness, interoperability, and cohesion.[52] While the NSO is primarily responsible for setting standards, other NATO entities—including in the NATO Science and Technology Organization (STO)—play important roles in implementing them and coordinating between national approaches.[53] Certification frameworks and the promulgation of best practices can similarly help incentivize the transposition of RRI into military organizations, even if standardization is by no means a purely military governance tool.

Both mechanisms, strategic policy planning and standards and certification, provide options for NATO to participate in AI governance regimes focusing on international security. NATO's operationalization of these tools may hold important implications while implementing successful AI governmental regimes for Allies and other defense stakeholders. In the next section, we consider each mechanism within foundational issues, or pillars, to illustrate NATO's role in AI governance.

NATO and Technological Change: Three Pillars of AI Governance

This section considers three pillars where NATO has procedures and competency to operationalize AI governance through both mechanisms of policy alignment and standards, and enhance security in the international environment. The pillars reflect foundational issue areas constitutive of governance but are also issue areas where previous scholars have cautioned as particularly challenging in the AI governance space. The three pillars—(1) ethics and values, (2) legal norms, and (3) safety and security—are meant to illustrate three conditions for NATO to facilitate policy and standards harmonization. Importantly, these pillars are not exhaustive areas in which NATO will need to consider governance structures to responsibly implement AI technology, but rather highlight particular issues that researchers and analysts acknowledge as significant hurdles in navigating AI governance (see Table 49.2).[54]

The first pillar considers NATO's role in the evolution of ethical and values-driven AI. One ongoing debate regarding AI as a military technology is the ethical implications and baseline values the Allies, and others, want infused in the development and adoption of AI. The Allies themselves lack uniform consensus on numerous, substantial ethical questions on the use of AI, as most clearly seen in the adjacent area of the ethics of autonomy in weapons systems including lethal autonomous weapon systems (LAWS). In this discussion, we spotlight NATO's role in facilitating and shaping ethical harmonization as an operational requirement to ensure successful future missions.

The second pillar examines legal norms as a domain wherein legal uncertainty regarding AI has tangible implications for Allied legal interoperability, a subset of larger coalition

Table 49.2 Cross-tabulating NATO's governance mechanisms with pillars of AI governance

	Ethics and Values	Legal Norms	Safety and Security
NATO Policy and Strategic Planning	Core, shared values at foundation of Alliance's political cohesion that informs civilian oversight of operations and overall institutional effectiveness; included in principles	Alignment between differing legal interpretations between Allies, particularly as affects the ability of forces to communicate and interoperate in dynamic contexts; constant calibration of policies based on legal interpretations	Strategic planning for maintaining integrity of information in military operations and transparency measures that reinforce democratic accountability; setting priorities on defensive systems and countermeasures to protect from motivated attacks and intentional failure modes of AI-enabled weapon systems
NATO Standardization and Certification	Basis of standards reinforcing responsible state behavior (see legal norms) and responsible engineering practices (see safety and security); human-centric views of responsibility and accountability also embedded in technical adoption measures	Precedents of legal standards dealing with international humanitarian law, including detention standards and training publications	Certification frameworks and technical, human, and procedural standards to prevent emergent behavior and enhance robustness and resilience of systems to behave predictably in conflict environments

interoperability. Thus far, the legal debate regarding AI has been largely fixed on the issue of a treaty banning the use of LAWS. In this section, we advocate for a more nuanced legal picture in which NATO can facilitate legal coordination and tackle some of the foundational legal issues which will prevent successful legal interoperability in future operations.

The third pillar identifies safety and security of AI systems as prerequisite to trustworthy and responsible AI in any context, but especially so for the conduct of military activity. At the NATO level, Allied forces must ensure their systems interoperate safely and predictably both to ensure effective command and control (C2) internally, and to prevent disruptions from attacks. It is a foundational facet of coordination that shows the overlap between NATO interests in military effectiveness and incentivization for responsible innovation.

Ethics and values

One of the vital aspects of AI which has garnered significant global attention is the ethical implications of artificial intelligence as a military technology—an issue that has divided

much of the global community, including NATO member states. As a starting point, researchers and analysts have considered the implications of emerging military technology in terms of ethical responsibility and regulation, especially as states and organizations continue to release AI ethical principles, guidelines, and standards.[55]

We explore how NATO can operationalize the debate around ethics and values of military AI to garner coordination and continue progress of EDT harmonization among partners. Building on the theoretical discussion from STS and military innovation literature above, the adoption of technologies that reinforce values serves the strategic interest of NATO to shape technological innovation against current waves of illiberalism.

Additionally, infusing AI development with certain ethical principles and values can have operational advantages and benefits, and NATO can, in particular, promote the ethical principles as operational standards for the Allies. A common critique within the ethics debate is that approaching new technology with an ethical or democratic values-driven perspective translates into comparative military disadvantage. Essentially, if your adversary develops technology without the constraints of ethical principles then there will be diminished effectiveness on the battlefield.[56] We find this critique unfounded because it assumes there is a false trade-off between ethics and effectiveness; instead, we argue ethical foundations are built into the architecture of modern warfare.[57] As such, ethics is a background condition for battlefield effectiveness, which is already infused in military decision-making and helping to guide the boundaries of international humanitarian law. As such, ethical guidelines do not have to detract from a military's capacity or competency to devise means and methods of warfare that will serve their national or coalition interest.[58] If anything, a first-mover advantage can incentivize an ethical and values-driven AI to establish the threshold of technological standards globally.[59]

The political dimension of the Alliance rests on the bedrock of a shared commitment to the "principles of democracy, individual liberty and the rule of law," as enshrined in the foundational North Atlantic Treaty of 1949.[60] Shared values are important for NATO operations because they help constitute their legitimacy. In addition to the North Atlantic Council exerting civilian oversight over NATO operations, legitimacy also includes respect for international legal principles including the core principles of international humanitarian law, or the laws of armed conflict, distinction, proportionality, and necessity. Without political oversight and legitimacy, NATO's military power would be less effective at shaping norms and promoting stability in the international system. The introduction of AI means that NATO has the moral and strategic imperative to adopt technologies that confer legitimacy and responsible innovation.[61] Acting on a shared commitment to democratic values is vital to the political cohesion of the NATO Alliance, just as much as it is a determinant of military effectiveness in a predictable security environment.

Put simply, shared values are important to both political and operational coherence between Allies. In its 2018 Framework for Future Alliance Operations, the strategic command Allied Command Transformation urged discussion of the legal and ethical dimensions of technological advancement to both know how it would impact NATO decision-making and how the Alliance would be prepared to address adversaries who do not share in that vision.[62] As such, NATO is contending with the ways that ethical AI impacts its own cohesion internally and how differences between allies may project outward in the face of competitors whose ethical frameworks and commitment to the rule of law differ. Internally, there is a strong national government commitment to responsible AI. Recently, transatlantic cooperation has initiated partnerships of largely NATO states committed to advancing responsible

AI with goals towards data sharing and future interoperability.[63] AI defense partnerships are not restricted to military innovation but rather aim to facilitate civilian innovation cooperation for defense purposes.

Externally, as AI-enabled autonomous systems enter the arsenals of more technologically advanced countries, uncoordinated ethical frameworks between Allies could pose operational risks. Without wider alignment, AI systems will have "varying technical specifications based on the legal and policy decisions made by individual governments when answering the key questions."[64] Further, although one motivation of autonomous systems is the increased safety of military personnel by removing them from dangerous situations, the lack of alignment could lead some Allies to perceive other capitals' deployments of unmanned forces as a lack of commitment to put lives on the line, therein posing credibility risks for Allies to assure one another.[65] These credibility risks can be mitigated by accountability and verification standards and procedures that NATO can implement for multinational operations, and efforts to institutionalize these procedures for AI are underway.[66] While the NATO AI Strategy is expected to create a common foundation for the Alliance's pursuit of AI, it is the implementation of principles for safe, ethical, legal, and interoperable AI that will reveal how coherent different national perspectives are. As of August 2021, only the United States and France have publicly issued their military AI strategies.[67] Other allies, including Canada and the United Kingdom, have emerging views on responsible military AI, but little official information about how they implement their ethical risk assessments is publicly available.[68]

NATO's influence in the functioning of joint operations and multinational military operations situates the Alliance to coordinate between how Allies implement ethical principles in their own national AI development. Specifically, NATO is well-situated to advocate for transparency, accountability, and data governance, which are also adoption factors that can translate into operational benefits, among other values.[69] For example, these factors can promote coordination among Allies on ethical guidelines on the development and use of AI, as this will be a necessary foundation in any future joint operation that uses this technology. "The transatlantic partnership must focus on coordinating these core principles and systematic governance to ensure AI systems development aligns with the rule of law and democracy. In particular, this must ensure answering questions about human dignity, human control, and accountability... NATO remains the organization that can bring these two (U.S. and EU) together and establishes the ethical bottom line."[70] The issues of transparency and accountability will define the scope of future implementation.

Many remaining questions and uncertainty will be addressed in NATO's forthcoming AI ethical principles guidelines. But the guidelines adopted in 2021 do not address every ethical dilemma. Regarding accountability, especially, likely major questions will continue to affect the Alliance. As Assistant Secretary-General for Emerging Security Challenges David van Weel recently clarified, NATO will offer a framework of responsible use for the Allies—but the question of accountability for member states, as opposed to civilian technology manufacturers for example, is one principle that will not have an easy solution.[71]

International legal norms

In certain respects, the legal debate mirrors much of the ethical debate surrounding AI as the two address many of the same issues. International law is a values-based system

embedded in certain principles and practices agreed-upon within the international system. This section certainly identifies the complementarity of the ethical discussion surrounding AI, but it also illustrates where the legal debate can depart from the ethical considerations to address different sorts of legal challenges that face the Alliance.

Lawyers, researchers, and civil society grapple with existing legal regimes relevant to military operations and the uncertainty and ambiguity surrounding automated decision-making, particularly in lethal decision-making. Thus far, the legal dialogue has been heavily anchored in the applicability of international humanitarian law (IHL), and other relevant legal regimes, to lethal autonomous weapons systems.[72] IHL, also known as the laws of war or the laws of armed conflict, regulate the means and methods of warfare and, as such, is pivotal to the emergence of military technology and how existing legal structures are disrupted. The legal debate often revolves around the prospect of a "treaty ban" of LAWS.[73]

But the legal debate is much more nuanced than the likelihood of international treaties banning any particular weapon system. Especially because NATO is not a regulatory body, it cannot institute measures to regulate emerging technology for the Allies. Instead, NATO's function in the legal domain may be more effective outside the traditional legal debates around emerging military technology and more embedded in fostering cooperation and coordination among military partners.

Other avenues of legal regulation may fall short of an international convention or prohibition, but nevertheless factor significantly in regulating and/or delineating state policies. Additionally, non-lethal applications of AI, as well as applications of AI that do not figure into autonomous systems, also raise important legal questions under international law. Arguably, norms around non-lethal applications are more urgent as their development is more advanced, harder to define, and less controversial in integration.[74] Ultimately, NATO's facilitative power can help ensure that integration of EDTs like AI into military capabilities and into multinational coalition operations is consistent with member states' legal obligations.

One vital and unique contribution for NATO is facilitating legal interoperability among the Allies to resolve some of the most pressing legal barriers for AI implementation in future Allied operations. Legal interoperability, a subset of larger coalition interoperability, refers to the operational coordination around partner legal obligations and interpretations.[75] It ensures "that within a military alliance, military operations can be conducted effectively consistent with the legal obligations of each nation."[76] Legal interoperability is a critical component of multilateral operations that has thus far been under-examined, despite its centrality to successful military operations. This is largely because "legal factors have a bearing on everything in alliances and coalition operations—from determining basic 'troop-to-task' considerations to decisions regarding the targets to be engaged—and the types of ordinances that may be used."[77]

To enhance legal interoperability, NATO can exert its influence on how Allies can develop and deploy AI consistent with their legal obligations through its unique standardization capacities. Historically, NATO has taken significant steps to bridge the legal gap between Allies on critical procedures that bridge responsible state behavior with such "troop-to-task" considerations. One instructive example from past operations is detention policies in non-international armed conflicts.[78] The promulgation of detention standards illustrates the operational significance of NATO's common legal procedures, even for coalitions of the willing that formally operate outside NATO structures. By way of background, the U.S.-led

coalition in Afghanistan had internal debates regarding the 96-hour security detention time period.[79] The United States advocated extending the 96-hour rule, where coalition partners insisted adhering to the NATO standard, even though it was not a NATO operation.[80] Generally the detention example illustrates NATO legal standards providing clarity to non-NATO operations; in some cases, Allies adopt NATO standards as accepted thresholds that continue to inform coalition policies beyond NATO structures and operations.

Implementing AI in future military operations will almost certainly complicate legal interoperability as there is a lack of uniform standards, as in the detention example. Even some of the more basic implementation measures will garner legal uncertainty and Allies will inevitably navigate with minimal legal clarity and no standard procedures. Despite the roots of the legal debate stemming from the question of lethality, the most pressing (and urgent) legal issues will address the integration of necessary AI-enablers, such as data gathering and sharing.

Furthermore, NATO has coordinated initiatives to promote awareness of Allies' legal obligations and has a dedicated office focusing on legality. This centralizes the institutional capacity to focus on alignment not only between the policies of NATO Allies, but coherence with the international community more broadly. Among others, the NATO Legal Practitioners' Workshop and inter-organizational dialogue between NATO, the UN, and the International Committee of the Red Cross (ICRC), the latter of which has a delegation to NATO that provides legal training and education to practitioners.[81] The NATO Office of Legal Affairs (OLA) itself can also play a central role in navigating the challenges to legal interoperability. As the example of detention standards illustrates, NATO has been successful in implementing legal standards which translated into operational clarity and coalition policy outside NATO operations.

As part of its focus on responsibility in its EDT agenda, NATO has opportunities to facilitate AI legal standard-setting and coalition policies to ensure safer and responsible use of AI in Allied operations.

Safety and security

For humans to meet ethical and legal commitments when developing and deploying AI, the systems themselves must be safe, secure, and reliable. More simply put, if humans and institutions interacting with AI do not have confidence that the systems will perform as expected, then they cannot assure that its development and deployment are responsible. This makes safety and security a key pillar of responsible AI governance for any actor.[82] As this section explores for NATO in particular, safety and security are indispensable to the Alliance's stated goals to focus its approach to EDTs in the areas of "deterrence and defense, capability development, legal and ethical norms, and arms control aspects."[83]

Politically, democratic militaries using AI cannot be accountable to their citizenries nor their coalition partners if they lack mechanisms to trace and explain how their systems are reliable. Accidents and interference with AI systems could likewise create political risks for the Alliance. For example, if deepfakes and micro-targeted information attacks compromise confidence in the integrity of information used to build a common operating picture, then the operational difficulties could also erode political trust between Allies in a few key ways. In the North Atlantic Council, disagreement about the integrity of information could slow

the decision-making body's ability to react to fast-changing operational realities.[84] Further, compromised AI systems may not only make it harder for forces to prevent harm to non-combatants, but also to prevent friendly fire. In this way, coalition forces arguably face even higher obligations to coordinate on the reliability of their systems, relative to adversaries and near-peer competitors that tend to operate alone. As such, responsible AI governance is not purely technical; policy alignment and strategic planning are likewise necessary to draw attention to risk management above the tactical level.

Even without being attacked, governability of AI in a NATO context also means understanding how AI-enabled and autonomous systems developed by the 30 Allies—and other partners—will interact with one another. NATO has expressed interest in governability as a principle of AI "to disengage or deactivate in case of unintended behavior,"[85] which echoes the U.S. Department of Defense definition of governable AI.[86] Disengaging adversaries is important to maintain de-escalation measures in conflict. For NATO, interoperability between systems also relates to governable AI because allies must also consider how the interactions between the 30 Allies' own AI-enabled and autonomous systems may result in unintended or emergent behavior.[87] This means that NATO has a responsibility to coordinate activities—be they technical exchanges, standardization efforts, or training and exercises—to build confidence that the systems perform as humans intend.[88] Without this coordination, the lack of interoperability of allied systems could lead to accidents, and separately, the potential loss of operational effectiveness also presents vulnerabilities for adversaries to exploit.

In addition to governability, NATO and its Allies are assessing the risks that bias, attacks, and lack of interpretability can introduce in relation to the anticipated uses of a given AI system.[89] In security and defense, new and heightened risks include poisoning of the information environment, deception systems and techniques, uncertainty about the performance of systems in new and unknown environments, and the possibility that tensions or accidents escalate at a faster tempo than humans and institutions can process, among others. These risks can manifest either in motivated attacks or unintentional failure modes.[90] In both cases, assuring and certifying that military assets are safe and secure is important given the inherently high risk in operational environments. These operational environments include the presumption that an adversary is disrupting one's own systems, be it by directly attacking the AI systems themselves, or disrupting the broader command, control, and communications systems under which the AI systems are operating.[91]

Mitigating these types of risks is typically done in testing, evaluation, validation and verification (TEVV) and in experimentation activities.[92] Yet AI cannot be validated and verified the way traditional software systems are because there is no guarantee that an AI system will perform in the real world as it does in a testing environment, and because lifelong-learning systems will perform differently over their lifecycle. Having robust assurance and TEVV processes in place are also important for operators to build trust in the systems they are meant to use, as well as for citizenries and coalition partners at large to see that accountability procedures still apply. As such, building institutional procedures to govern AI safety and security is necessary to build trust in the use of the technology—as well as to develop countermeasures and defensive systems that protect against adversarial threats.

NATO thus has an institutional responsibility to prevent and mitigate these intentional and unintentional failures if using AI in operations and mission support.[93] As Table 49.1 shows, the Alliance also has a range of relevant entities to coordinate national approaches

to AI safety and security, as well as facilitate safety measures as part of responsible use in the Alliance-wide ecosystem.

NATO has an important role to play in military standardization and Allied policy planning for safe, secure, and interoperable AI. This includes the coordinating role of the Conference of National Armaments Directors and the Command, Control and Consultation Board to implement complementary acquisition processes that fuse AI adoption measures with safety responsibilities. Furthermore, entities including STO and NSO have a significant role setting the technical baseline and promulgating materiel standards that provide the technical framework for safety and security. Although their staffs are themselves small, they both convene hundreds, if not thousands, of subject matter experts in working groups. As such they both offer unique technical networks to help shape safety and security in a way that minimize risk in operations. NATO's resources and leadership are vital to using standards and coalition policy to instill safe and secure technological development, a necessary condition to interoperable and successful future operations.

Conclusion

At the core, this chapter argues that NATO is well positioned to steward the development of military AI and institute governance mechanisms towards coalition inclusion of responsible AI while simultaneously maintaining incentives for comparative advantage. Using the three pillars—ethics and values, legal norms, and safety and security—as issue areas which present AI governance challenges, we show that NATO has space to emerge as a leader in AI governance and contribute to responsible adoption of EDTs in the international security environment. This builds on foundations that derive NATO's responsibilities to govern AI according to its values, legal obligations, and institutional interests. These foundations from both STS and military innovation studies offer ways that the Alliance can activate its existing governance mechanisms to exert influence in new ways. Not only is this influence important for the Alliance to bolster its institutional relevance in an evolving international security architecture, but it also dovetails with its capacity to shore up military effectiveness and interoperability as Allies modernize their arsenals and associated concepts into the frontier of AI.

Importantly, we do not argue NATO is the only—or even the most important—actor shaping AI governance in international security. Other contributors in this *Handbook* impressively detail efforts at both state and regional levels. Our aim has been to convince sceptics that NATO has a role that is not replicated by other stakeholders in the international security environment. NATO has particular influence, procedures, and the competency to institute certain governance mechanisms—namely standardization and policy planning—that it can build on without needing to expend time building new institutions from scratch. Beyond just a role, NATO is incentivized to emerge as a steward of AI governance and use these mechanisms for future operations, should the Alliance wish to maintain its unique position as a leader encouraging policy alignment, defense planning, and military standardization.

More broadly, this chapter illustrates that regional and international organizations have high stakes for military AI governance. As development, procurement, and implementation of AI is accelerating, it is imperative that international organizations facilitate cooperation among states and industry partners to guide responsible military AI implementation aligned with core values and legal obligations. The convening and coordinating power of international organizations, among other governance tools, is a necessary step for state cooperation and policy alignment. How exactly NATO interacts with other international organizations in the security architecture, including the UN and EU, is a political topic that will also have important implications for the composition of international technology governance regimes, and is a subject for further research.

On that note NATO, or any other international organization, is not exempt from these political hurdles. As EDTs increasingly become a focal point in the geopolitical space, any approach of AI governance in the international security environment will have global political undertones. This will undoubtedly be a significant hurdle for NATO as it balances responsible AI development and Allied coordination and cooperation in a changing geopolitical landscape. And certainly, the political realities may well represent the greatest challenge and disincentivize NATO to emerge as a leader in responsible military AI. Nevertheless, the three pillars indicate that NATO is an institution with considerable opportunity to shape responsible AI governance. More specifically, this entails urging and facilitating Allied standards and policies to establish foundations for emerging military technology built on informed and ethical principles and enhance the international security environment.

In any discussion of AI as an emerging military technology, it is necessary to strike a balance between acknowledging the transformative potential of AI in the security environment, while simultaneously recognizing the "hype" that may, thus far, be unfounded. But some conclusions are clear. The risks and opportunities of military AI can pose significant challenges for future military operations, and this necessarily means there are many stakeholders with a vested interest in developing, promoting, and implementing responsible military AI. As multinational coalitions and military operations are a foundational security policy for much of the world, this means NATO is also a stakeholder with a vital interest in promoting safe and secure technology among its partners, both traditional and non-traditional. As the international security environment continues to shift, there is space for NATO to pursue its agenda to maintain technological superiority not just to protect and defend its way of life, but also to build on its pillars of AI governance to steward military innovation on a responsible trajectory.

Acknowledgments

Zoe Stanley-Lockman was previously an Associate Research Fellow at the Institute of Defence and Strategic Studies at the S. Rajaratnam School of International Studies (RSIS) and contributed to this chapter in a personal capacity. Lena Trabucco is a Postdoctoral Researcher at the Centre for Military Studies at the University of Copenhagen. Both authors contributed equally to this chapter. The authors wish to thank Matthijs Maas, Joanna van der Merwe, and Simona Soare for their helpful comments.

Notes

1. The three core tasks, as set out in the 2010 Strategic Concept, are collective defense, crisis management, and cooperative security. NATO is currently updating its Strategic Concept to be prepared for adoption in 2022. It has announced it is updating its Strategic Concept in the near future given significant changes to the strategic environment in intervening years, including. The forthcoming concept will contend with changes to the strategic environment, including adaptation toward increasing technology-related threats and opportunities, as well as the rise of China.
2. Gilli, Andrea. (2021, February). "NATO, technological superiority, and emerging and disruptive technologies." In Thierry Tardy (Ed.), *NATO 2030: new technologies, new conflicts, new partnerships* (p. 6). NATO Defense College, Rome.
3. Sprenger, Sebastien. (2020, October 9). NATO chief seeks technology gains in alliance reform push. *Defense News*.
4. The Economist. (2019, November 7). Emmanuel Macron warns Europe: NATO is becoming brain-dead." *The Economist*. https://www.economist.com/europe/2019/11/07/emmanuel-macron-warns-europe-nato-is-becoming-brain-dead.
5. See for example, Williams, Michael John. (2013). Enduring, but irrelevant? Britain, NATO and the future of the Atlantic alliance. *International Politics 50*; Hellmann, Gunther, & Wolf, Reinhard. (1993). Neorealism, neoliberal institutionalism, and the future of NATO. *Security Studies 3* (1).
6. For more on AI as a disruptive technology, see Liu, Hin-Yan, et al. (2020). Artificial intelligence and legal disruption: A new model for analysis. *Law, Innovation and Technology 12* (2); Sprenger, Sebastien. (2020, October 9). NATO chief seeks technology gains in alliance reform push. *Defense News*. https://www.defensenews.com/global/europe/2020/10/09/nato-chief-seeks-technology-gains-in-alliance-reform-push/.
7. There is a host of literature that looks at the ways that AI enables and scales up risks in the security and defense environment, including in the field of existential risk. Other chapters of this *Handbook* introduce how AI is an instrument of national power and what concomitant risks this introduces into international politics. For more information on AI-related risks in security and defense, see Arsenault & Kreps. (2024). AI and international politics. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI governance*. Oxford University Press; Horowitz, Pindyck, & Mahoney. (2024). AI, the International Balance of Power, and National Security Strategy. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI governance*. Oxford University Press.
8. Lazar, Seth. (2024). Power and AI: Nature and justification. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek, & V. Hudson (Eds.), *The Oxford Handbook of AI governance*. Oxford University Press.
9. Biermann, Frank, et al. (2009). The fragmentation of global governance architectures: A framework for analysis. *Global Environmental Politics 9* (4), 15.
10. Further, while beyond the scope of this chapter, it is worth noting that the OECD plays an important role in the international technology governance regime. The OECD AI Principles have been adopted by several transatlantic countries, including the United States, and Franco-Canadian leadership in both the OECD and the G7 have spurred important initiatives on AI governance (including the establishment of the Global

Partnership on AI, or GPAI). Influences of these other stakeholders on NATO's adoption and governance of AI may ultimately be similar to the way it imports characteristics of nation-state and EU approaches to governance, but they are not discussed at length because they do not focus on defense. See Franke, Ulrike. (2021, June 21). Artificial intelligence diplomacy: Artificial intelligence governance as a new European Union external policy tool. *European Parliament Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies*. https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU%282021%29662926; Boulanin, Vincent, Brockmann, Kolja, & Richards, Luke. (2020, November). Responsible artificial intelligence research and innovation for international peace and security. *Stockholm International Peace Research Institute*. https://www.sipri.org/publications/2020/other-publications/responsible-artificial-intelligence-research-and-innovation-international-peace-and-security.

11. A notable exception is Gilli, Andrea. (2020). "NATO-Mation" strategies for leading in the age of artificial intelligence. *NATO Defense College Research Paper 15*.
12. Use is one, but not the only, phase of adoption. We intentionally focus on use here to echo NATO's own language on responsible use, understanding that the Alliance has more authority in encouraging and coordinating activities than it does in development. Regulation is an important mechanism that is beyond the scope of this chapter because NATO is not a regulatory body.
13. Drezner, Daniel W. (2019). Technological change and international relations. *International Relations 22* (2); Horowitz, Michael. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review 1* (3); Milner, Helen. (2006). The digital divide: The role of political institutions in technology diffusion. *Comparative Political Studies 39* (2); Drezner, Daniel W. (2004). The global governance of the internet: Bringing the state back in. *Political Science Quarterly 119* (3).
14. Roland, Alex. (2016). *War and technology: A very short introduction* (p. 5). Oxford University Press.
15. Dafoe, Allan. (2015). On technological determinism: A typology, scope and conditions, and a mechanism. *Science, Technology & Human Values 40* (6).
16. MacKenzie, Donald, & Wajcman, Judy. (1999). *The social shaping of technology*. Open University Press; Stilgoe, Jack, Owen, Richard, & Macnaghten, Phil. (2013). Developing a framework for responsible innovation. *Research Policy 42*, 1568–1580; Arthur, Brian W. (2009). *The nature of technology: What it is and how it evolves*. Penguin Books.
17. Dignum, Virginia, Muller, Catelijne, & Theodorou, Andreas. (2020, February). First analysis of the EU whitepaper on AI. *ALLAI*. http://allai.nl/first-analysis-of-the-eu-whitepaper-on-ai/ and Hwang, Tim. (2018, March 23). Computational power and the social impact of artificial intelligence. *SSRN*, p. 2. https://ssrn.com/abstract=3147971.
18. Baxter, Gordon, & Sommerville, Ian. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers 23* (1), 4–17.
19. Boulanin et al. (2020); Genus, Audley, & Stirling, Andy. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research Policy 47* (1), 61–69; see also Verbruggen, Maaike. (2019). The role of civilian innovation in the development of lethal autonomous weapon systems. *Global Policy 10* (3).
20. Owen, Richard, Stilgoe, Jack, Macnaghten, Phil, Gorman, Mike, Fisher, Erik, & Guston, Dave. (2013). A framework for responsible innovation. In Maggy Heintz, J. Bessant, &

Richard Owen (Eds.), *Responsible Innovation: Managing the responsible emergence of science and innovation in society* (p. 36). John Wiley & Sons.

21. Military adoption of commercially driven technologies is sometimes referred to as a "spin on" process. Here, the legal, societal, and cultural elements that impact the development of a dual-use technology would similarly spin on, or be imported, to the military adopter. For more on the term "spin on," see Stowsky, Jay. (1991, May). From spin-off to spin-on: Redefining the military's role in technology development. Working Paper.
22. Stilgoe, Jack, Owen, Richard, & Macnaghten, Phil. (2013). Developing a framework for responsible innovation. *Research Policy 42*, 1572.
23. Collingridge wrote: "When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult and time consuming." See Collingridge, David. (1980). *The social control of technology* (p. 11). St. Martin's Press.
24. Owen et al. (2013), 35.
25. Collingridge (1980).
26. Stanley-Lockman, Zoe. (2021, March 10). Emerging AI governance for international security: The stakes and stakeholders of responsibility. *Azure Forum*. https://www.azureforum.org/the-azure-forum-strategic-insights/emerging-ai-governance-for-international-security-the-stakes-and-stakeholders-of-responsibility/.
27. Dafoe (2015), 1048–1049.
28. Case studies typically focus on military services or non-state insurgents, rather than supra-national organizations, in the literature on adaptation of doctrine, structures, tactics, techniques, and procedures.
29. Horowitz, Michael. (2010). *The diffusion of military power: Causes and consequences for international politics*. Princeton University Press; Kier, Elizabeth. (1997). *Imagining war: French and British military doctrine between the wars*. Princeton University Press.
30. Shimshoni, Jonathan. (1990–1991). Technology, military advantage, and World War I: A case for military entrepreneurship. *International Security 15* (3), 187–215.
31. Mahnken, Thomas. (2008). *Technology and the American way of war since 1945* (p. 219). Columbia University Press.
32. Shimshoni (1990–1991), 189.
33. For more on AI governance more broadly, see Cihon, Peter, Maas, Matthijs M., & Kemp, Luke. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy 11* (5).
34. Citing Barnett, Correlli. (1963). *The swordbearers: Studies in supreme command in the First World War* (p. 11). Eyre & Spottiswoode. Grissom, Adam. (2006). The future of military innovation studies. *Journal of Strategic Studies 29* (5), 907.
35. Correlli (1963), 11; Grissom (2006), 907.
36. Mahnken (2008), 2.
37. Raudzens, George. (1990). War-winning weapons: The measurement of technological determinism in military history. *Journal of Military History 54* (4), 403–33; Phillips, Gervase. (2002). The obsolescence of the *Arme Blanche* and technological determinism in British military history. *War in History 9* (1), 39–59; Roland, Alex. (2010). Was the nuclear arms race deterministic? *Technology and Culture 51* (2), 444–461; Rogers, Clifford J. (2011). The development of the longbow in late medieval England and "technological determinism". *Journal of Medieval History 37* (3), 321–341; Pavelec, Sterling Michael. (2012). The inevitability of the weaponization of space: Technological constructivism versus determinism. *Astropolitics 10* (1), 39–48; Kuo, Kendrick. (2021). Military innovation and technological

determinism: British and U.S. ways of carrier warfare, 1919–1945. *Journal of Global Security Studies 6* (3), 1–19.

38. Hughes, Thomas. (1987). The evolution of large technological systems. In W. Bijker, T. Hughes, & T. Pinch (Eds.), *The Social Construction of Technological Systems*. MIT Press; Fino, Steven. (2015, January.). All the missiles work: Technological dislocations and military innovation: A case study in U.S. Air Force air-to-air armament post-World War II through Operation Rolling Thunder. Air University School of Advanced Air and Space Studies, 22–23.
39. Fino (2015), 41.
40. Dafoe (2015), 1049.
41. Fino (2015), 45.
42. Dafoe (2015), 1069.
43. The authors thank Maaike Verbruggen for pointing us toward the NATO Advanced Research workshop on "Social Responses to Large Technical Systems: Regulation, Management, or Anticipation" that took place in Berkeley, California from October 17–21, 1989 as one example of this engagement.
44. Examples include the socio-technical "NATO Human View" framework as part of the 2010 technical report "Human Systems Integration for Network Centric Warfare," as well as the "Futures Assessed alongside socio-Technical Evolutions (FATE) Method" developed from 2018–2021.
45. Leonardi, Paul M., & Barley, Stephen R. (2010). What's under construction here? Social action, materiality, and power in constructivist studies of technology and organizing. *The Academy of Management Annals 4* (1), 3.
46. Christie, Edward Hunter. (2020, November 24). Artificial intelligence at NATO: Dynamic adoption, responsible use. *NATO Review*. https://www.nato.int/docu/review/articles/2020/11/24/artificial-intelligence-at-nato-dynamic-adoption-responsible-use/index.html.
47. The Organisation for Economic Co-operation and Development (OECD) has identified 10 instruments in its work on the "ethics of emerging technologies" policy theme, which offers a broader list of possible technology governance mechanisms. They are regulatory oversight and ethical advice bodies, formal consultation of stakeholders or experts, national strategies/agendas/plans, emerging technology regulation, policy intelligence (e.g., evaluations, benchmarking, and forecasts), networking and collaborative platforms, creation or reform of governance structure or public body, standards and certification for technology development and adoption, information services and access to datasets, and grants for business R&D and innovation. See Organisation for Economic Co-operation and Development. (2021). Technology governance (accessed April 29, 2021). https://www.oecd.org/sti/science-technology-innovation-outlook/technology-governance/.
48. Deni, John R. (2020). *Security threats, American pressure, and the role of key personnel: How NATO's defence planning process is alleviating burden sharing*. U.S. Army War College Press.
49. Lawrence, Christie, & Cordey, Sean. (2020, August). The case for increased transatlantic cooperation on artificial intelligence. *Harvard Kennedy School Belfer Center for Science and International Affairs*, 85. https://www.belfercenter.org/publication/case-increased-transatlantic-cooperation-artificial-intelligence.
50. NATO. (2020, November). *NATO 2030: United for a New Era: Analysis and Recommendations of the Reflection Group Appointed by the NATO Secretary General*, p. 29. https://www.nato.int/nato_static_fl2014/assets/pdf/2020/12/pdf/201201-Reflection-Group-Final-Report-Uni.pdf.

51. NATO Standardization Office. (2004, January 13). NATOTerm (definition). https://nso.nato.int/natoterm/Web.mvc.
52. Beckley, Paul. (2020, July). Revitalizing NATO's once robust standardization programme. *NATO Defense College*, 3–4.
53. See, for instance, the NATO Modelling and Simulations Group within STO focuses on standards related to testing and experimentation, among other functions. See Stanley-Lockman, Zoe. (2021, August). Military AI cooperation toolbox: Modernizing defense science and technology partnerships for the digital age. *Center for Security and Emerging Technology*, 36–37.
54. For example, arms control is a vital pillar of AI governance that we do not address explicitly in this chapter.
55. Jobin, Anna, Ienca, Marcello, & Vayena, Effy. (2019, September). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*, 389–399.
56. This argument about ethical constraints putting democratic forces at a competitive disadvantage is usually part of a broader framing of an AI arms race, and often specifically focuses on autonomy in weapons rather than just AI ethics. See Boudreaux, Benjamin. (2019, January 11). Does the U.S. face an AI ethics gap? *Real Clear Defense*. https://www.realcleardefense.com/articles/2019/01/11/does_the_us_face_an_ai_ethics_gap_114095.html; Thornton, Rod. (2019, June 17). One to ponder: The UK's ethical stance on the use of Artificial Intelligence in weapons systems. *Defence in Depth*. https://defenceindepth.co/2019/06/17/one-to-ponder-the-uks-ethical-stance-on-the-use-of-artificial-intelligence-in-weapons-systems/; Roper, Will (2020, October 24). There's no turning back on AI in the military. *Wired*. https://www.wired.com/story/opinion-theres-no-turning-back-on-ai-in-the-military/; Morgan, Forrest E., Boudreaux, Benjamin, Lohn, Andrew J., Ashby, Mark, Curriden, Christian, Klima, Kelly, & Grossman, Derek. (2020). *Military Applications of Artificial Intelligence Ethical Concerns in an Uncertain World, RAND Corporation*, 2(11–12), 41. https://www.rand.org/pubs/research_reports/RR3139-1.html.
57. The National Security Committee on Artificial Intelligence (NSCAI) acknowledged this point in their 2019 report, "Everyone desires safe, robust, and reliable AI systems free of unwanted bias, and recognizes today's technical limitations. Everyone wants to establish thresholds for testing and deploying AI systems worthy of human trust and to ensure that humans remain responsible for the outcomes of their use. Some disagreements will remain, but the Commission is concerned that debate will paralyze AI development. Seen through the lens of national security concerns, inaction on AI development raises as many ethical challenges as AI deployment. There is an ethical imperative to accelerate the field of safe, reliable, and secure AI systems that can be demonstrated to protect the American people, minimize operational dangers to U.S. service members, and make warfare more discriminating, which could reduce civilian casualties." National Security Commission on Artificial Intelligence. (2019, November). *Interim Report*. Washington, D.C., 16–17.
58. We thank Kate Devitt for raising this point.
59. See Horowitz (2018) for more on AI and first-mover advantage. See also Liivoja, R., & McCormack, T., (Eds). (2016). *Routledge Handbook of the Law of Armed Conflict*. Routledge.
60. North Atlantic Treaty (April 4, 1949): "*founded on the principles of democracy, individual liberty and the rule of law" and "promoting conditions of stability and well-being*."

61. Gilli, Andrea, & Stanley-Lockman, Zoe. (2020). Ethical purpose: Ethics and values. In *NATO-Mation: Strategies for Leading in the Age of Artificial Intelligence* (pp. 29–30). NDC Research Paper.
62. Allied Command Transformation. (2018). Framework for future Alliance operations, 15–6.
63. https://www.ai.mil/news_09_16_20-jaic_facilitates_first-ever_international_ai_dialogue_for_defense_.html; see also Trabucco, Lena. (2020). AI partnership for defense is a step in the right direction—but will face challenges. http://opiniojuris.org/2020/10/05/ai-partnership-for-defense-is-a-step-in-the-right-direction-but-will-face-challenges/.
64. van der Merwe, Joanna. (2021). Establishing NATO ethical AI principles is the first step toward both technical and political alignment. *CEPA*. https://cepa.org/nato-leadership-on-ethical-ai-is-key-to-future-interoperability/.
65. Wong, Yuna Huh, Yurchak, John M., et al. (2020). Deterrence in the age of thinking machines. *RAND Corporation 6*, 60.
66. For more on governance tools such as accountability and verification procedures, see Salamon, Lester. (2002). *The tools of government: A guide to the new governance*. Oxford University Press.
67. Hill, Steven. (2020). AI's impact on multilateral military cooperation: Experience from NATO. Presented at *Symposium: How will Artificial Intelligence Affect International Law?* (p. 148); Stanley-Lockman, Zoe. (2021, August). Responsible and ethical military AI: Allies and Allied perspectives. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/.
68. Hill (2020), 148; Stanley-Lockman (2021).
69. See for example Gilli & Stanley-Lockman (2020); Gilli, A., Pellegrino, M., & Kelly, R. (2019). Intelligence machines and the growing importance of ethics. In A. Gilli (Ed.), *The Brain and the processor* (pp. 45–54). NATO; Tiell, S. C., & Mertcalf, B. (2016). *Universal principles of data ethics: 12 guidelines for developing ethics codes*. Accenture; Roa, A., Palaci, F., & Chow, W. (2019). *A practical guide to responsible artificial intelligence* PricewaterhouseCoopers.
70. van der Merwe (2021).
71. Sprenger, Sebastian. (2021, April 27). NATO tees up negotiations on artificial intelligence in weapons. *C4ISRnet*. https://www.c4isrnet.com/artificial-intelligence/2021/04/27/nato-tees-up-negotiations-on-artificial-intelligence-in-weapons/.
72. While the debate has been heavily focused on IHL applicability issues, there is significant uncertainty surrounding international human rights law and domestic legal obligations. See for example, Donahoe, Eileen, & Metzger, Megan MacDuffee. (2019). Artificial intelligence and human rights. *Journal of Democracy 30* (2); Raso, Filippo, Hilligoss, Hannah, Krishnamurthy, Vivek, Bavitz, Christopher, & Kim, Levin Yerin. (2018, September 25). *Artificial intelligence & human rights: Opportunities & risks*. Berkman Klein Center Research Publication No. 2018-6. https://ssrn.com/abstract=3259344 and http://dx.doi.org/10.2139/ssrn.3259344.
73. Proponents of an international treaty that prohibits LAWS argue that a multilateral ban would effectively remove ambiguity and uncertainty around AI in lethality questions and set clear guidelines and expectations to the limits of autonomous weapon systems. However, skeptics of an international treaty ban point to the political hurdles inherent to any international treaty, let alone one for technology that is difficult to define, verify, and enforce. Importantly, skeptics are not necessarily against the regulation of autonomous lethal decision-making but believe that a blanket ban could also preclude positive

outcomes associated with technology adoption, including the reduction of human error. For academic commentary, see Garcia, D. (2015). Killer robots: Why the U.S. should lead the ban. *Global Policy 6* (1); Crootof, Rebecca. (2014–2015). The killer robots are here: Legal and policy implications. *Cardozo Law Review 36*; Goose, Stephen, & Wareham, Mary. (2016). The growing international movement against killer robots. *Harvard International Review 37* (4).

74. For instance, the International Committee of the Red Cross position paper on AI and machine learning in armed conflict identifies three main categories for conflict parties to use AI that pose risk from a humanitarian perspective. They are increasing autonomy, new means of cyber and information warfare, and the changing nature of decision-making in conflict. See International Committee of the Red Cross. (2021, March 1). Artificial Intelligence and machine learning in armed conflict: A human-centered approach. https://international-review.icrc.org/articles/ai-and-machine-learning-in-armed-conflict-a-human-centred-approach-913.
75. Coalition interoperability refers to, "the ability to act together coherently, effectively and efficiently, to achieve Allied tactical, operational, and strategic objectives" https://nso.nato.int/natoterm/Web.mvc. Another definition from the North Atlantic Treaty Organization, *Allied Joint Doctrine for Air and Space Operation*, P 2.2.1, AJP-3.3 Ed. B Version 1 (April 2016) describes interoperability as, "the effectiveness of Allied forces in peace, crisis or conflict, depends on the ability of forces provided to operate together coherently effectively and efficiently."
76. Hill, Steven, & Lemétayer, David. (2016). Legal issues of multinational military operations: An Alliance perspective. *Military Law and the Law of War Review 55*, 13.
77. Kelly, Col. Michael. (2005). Legal factors in military planning for coalition warfare and military interoperability: Some implications of the Australian Defence Force. *Australian Army Journal 2* (2), 690.
78. Hill, Steven, & Holzer, Leonard. (2019). Detention operations in non-international armed conflicts between international humanitarian law, human rights law and national standards: A NATO perspective. *Israel Yearbook on Human Rights 49*.
79. For more on legal issues regarding NATO's detention standards, see NATO Legal Gazette. (2019, May). *Significant Issues for the NATO Legal Community 39*. https://www.act.nato.int/application/files/7616/0999/3871/legal_gazette_39.pdf. This illustration is not to minimize the legal uncertainties that did exist for the detention regime. Legal complexities such as the legal basis for detention and the status of the armed conflict were central legal issues for the coalition. Rather, to illustrate the NATO standard helped provide a standard procedure for implementing a detention regime, even when faced with legal uncertainty about the conflict.
80. Trabucco, Lena. (2020). Judges on the Battlefield? Judicial Observer Effects in US and UK National Security Policies. PhD Diss. Northwestern University (p. 257).
81. Hill, Steven, & Lemétayer, David. (2016). Legal issues of multinational military operations: An Alliance perspective. *Military Law and the Law of War Review 55*, 13.
82. For a categorization of AI safety problems that has been influential in the field, see Amodei, Dario, Olah, Chris, Steinhardt, Jacob, Christiano, Paul, Schulman, John, & Mané, Dan. (2016). Concrete problems in AI safety. *arXiv preprint, arxiv:1606.06565*. https://arxiv.org/abs/1606.06565.

83. North Atlantic Treaty Organization. (2019, November 29). NATO: Ready for the Future adapting the Alliance (2018–2019), p. 17. https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2019_11/20191129_191129-adaptation_2018_2019_en.pdf.
84. Valášek, Tomáš. (2017, August 31). How artificial intelligence could disrupt alliances. *Carnegie Europe*. https://carnegieeurope.eu/strategiceurope/72966.
85. Christie (2020).
86. According to the United States, "DoD AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior." See Defense Innovation Board. (2019, October 31). AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense: Supporting document, p. 38.
87. Reding and Eaton (2020), 66; Konaev, Margarita, Chahal, Husanjot, Fedasiuk, Ryan, Huang, Tina, & Rahkovsky, Ilya. (2020, October). U.S. military investments in autonomy and AI. *Center for Security and Emerging Technology*, p. 15. See also Soare, Simona. (2021). What if... the military AI of NATO and EU states is not interoperable? In Flourence Gaub (Ed.), *What if... not? The cost of inaction*. European Union Institute for Security Studies.
88. This assumes that the humans designing the system accurately transpose their intended outcome with the reward signals, which is not necessarily a given, but is also beyond the NATO remit as these are national decisions.
89. Stanley-Lockman, Zoe. (in press). Responsible military AI: Allies and allied perspectives. *Center for Security and Emerging Technology*.
90. Technical failure can occur unintentionally, such as when performance is technically correct but produces unsafe consequences, or when ethics and safety considerations are not sufficiently built into the design phase and sustained in the life cycle of the system. They can also occur intentionally, when motivated actors attack the system to "misclassify the result, infer private training data, or to steal the underlying algorithm." See Shankar, Ram, Kumar, Siva, O'Brien, David, Snover, Jeffrey, Albert, Kendra, & Viljoen, Salome. (2019, November 11). Failure modes in machine learning. *Microsoft Corporation*. https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning.
91. French Ministry of Armed Forces. (2019, September). *Artificial intelligence in support of defence*, p. 9.
92. See for example, Flournoy, Michèle A., Haines, Avril, & Chefitz, Gabrielle. (2020, October). *Building trust through testing: Adapting DOD's test & evaluation, validation & verification (TEVV) enterprise for machine learning systems, including deep learning systems*. WestExec Advisors.
93. Christie (2020).

Index

Tables, figures, and boxes are indicated by *t*, *f*, and *b* following the page number

A
AAA (Algorithmic Accountability Act of 2019), 413, 503–504, 578–579
AAAI (Association for the Advancement of Artificial Intelligence), 885
Abraham, David, 991
ACAT (Appeals Council Analysis Tool), 783, 786, 788
accident models, 444–445
accountability, 164–176
  of agents, 82n4, 165–167, 169
  AI as instrument of, 174–175
  algorithmic bias and, 164, 525, 846–847
  audits for, 496, 497, 499–505, 500*f*, 510
  automated decision-making and, 465–466
  black box technologies and, 173, 969
  constructivist approach to, 487
  for decision-making, 54, 58, 168
  democratic, 167, 175, 176, 998–999
  for environmental degradation, 991
  ethics and, 165, 169–170, 173–174, 176, 221
  of EU's trustworthy AI, 943
  forensic, 165–167
  gaps in, 164, 167–170
  incentive-based rating systems and, 519, 525
  lethal autonomous weapons and, 897–898, 960
  mechanisms of, 167, 170–174
  in multinational operations, 1028
  permit programs and, 300, 307
  power and, 499
  in public AI systems, 422, 423, 428, 430–434
  safety issues and, 453
  of supply chains, 266
  surveillance AI and, 810
  tradeoffs involving, 384
  as virtue, 165, 166
accuracy
  of decision-making, 140, 464, 485
  deep learning processes and, 918
  of diagnostic tools, 839–840
  explainable AI and, 189, 190
  of facial recognition technology, 240, 385, 507, 508, 825
  of predictive algorithms, 133, 135, 137, 464
  question exploration tool on, 116
  reinforcement learning and, 964
  of SSA Disability Program decisions, 782
  of treaties, 974
Acemoglu, Daron, 14, 92, 102n16, 103n23, 642, 645, 647, 649, 650, 653–654, 656n3, 660, 664, 666, 668–673, 677–681, 684, 687–688, 691–692, 694–695, 697, 701, 729–730, 732–733, 750, 921
Ackerman, Bruce, 719, 722n19
ACLU (American Civil Liberties Union), 502, 508, 510, 998
administrative behavior, 383, 386–388, 390–393
administrative evil, 79, 393, 623
Administrative Law Judges (ALJs), 781, 784, 788
advertisements
  algorithmic bias and, 132, 520
  beauty standards in, 220
  manipulative, 50
  persuasive technologies and, 204
  political, 95–96, 239
  racial bias and, 132, 520
  targeted, 96, 141, 154, 203, 467–468, 523, 560
  transparency of, 262–263
affective computing, 109–115, 124, 125n1, 125n9, 126n6, 521
affirmative action, 139, 144n27, 218

AFST (Allegheny Family Screening Tool), 844
Agarwal, P. K., 388
agents and agency
  accountability of, 82n4, 165–167, 169
  bounded rationality of, 383, 386, 390–391, 394
  constraints on, 78–79
  of consumers, 519, 529
  distributed agency, 168–169
  of Global South, 986
  goals and, 22, 66–69, 68*f*, 82–83n9
  moral agents, 82n4, 152
  in political communities, 233, 234, 246
  power of, 200–202
  principal–agent relationships, 30, 67–69, 68*f*, 82–83n9, 166, 731
  in public organizations, 383–385, 390–392
  of robotic systems, 67
  technological developments and, 1020
Aggarwal, Nakul, 15–16, 838
Aggarwal, Nikita, 16, 860
Aguirre, Anthony, 8–9, 320
Ahn, Michael, 9, 10, 421
AI. *See* artificial intelligence
AIA. *See* Artificial Intelligence Act of 2021
AIAs. *See* algorithmic impact assessments
AIDM (Artificial Intelligence Decision-Making). *See* automated decision-making
AI ecosystem, 398–415
  audits of, 290
  complexity of, 401
  evolution of, 1–3, 2*f*, 399
  global nature of, 884, 888, 929
  governance of, 402–415, 402*f*, 404–406*f*, 409–411*f*, 414*f*
  informal norms within, 78
  long-term impacts, 425
  strengthening, 947–951
  use of term, 400–401
AI effect, defined, 556
AI governance
  accountability in (*see* accountability)
  application-centric approaches, 360, 361
  audits and (*see* audits)
  centralized, 89, 575, 576
  challenges of, 2–3, 32, 624–627
  convening on (*see* multi-stakeholder convening)
  corporate, 30, 732
  definitions of, 1, 21, 31, 403, 573, 1016
  domestic, 28, 33–34, 963–966, 969
  domestic conflicts and, 33–34
  ecosystem framework for, 402–415, 402*f*, 404–406*f*, 409–411*f*, 414*f*
  ethical issues (*see* ethics)
  in EU (*see* European Union)
  explainable AI and, 183–195
  fairness in (*see* fairness)
  in Global South (*see* Global South)
  goals of, 3, 45, 55–58, 75, 192–194
  great power competition and, 34–39
  holistic sensibility of, 24
  horizontal regulation, 937–938, 953
  inclusivity in (*see* inclusivity)
  institutional fit and, 31–32
  justice in (*see* justice)
  law-centric approaches, 360–361
  lifecycle approach to, 424–425, 428, 843
  by NATO, 1016–1017, 1021–1032, 1023–1024*t*, 1026*t*
  networked, 575–577, 580–581
  permits and (*see* permit programs)
  privacy issues (*see* privacy)
  for public organizations (*see* public organizations)
  scalability of, 23, 372–373
  social alignment and, 66, 71
  structural injustice in, 217–219
  technical metrics for, 345–356, 347*f*, 350–351*f*
  technology-centric approaches, 360, 364
  traditional models of, 219–223, 400, 573–576
  transnational (*see* transnational AI governance)
  transparency in (*see* transparency)
  vertical regulation, 629
AI HLEG (High-Level Expert Group on Artificial Intelligence), 848, 940–943, 945
Aiken, Howard, 100n2
AI Partnership for Defense (PfD), 928
AJOPT (Algorithmic Justice and Online Platform Transparency) Act (proposed), 572–573, 579
Ajunwa, Ifeoma, 505

Algorithmic Accountability Act of 2019 (AAA), 413, 503–504, 578–579
algorithmic bias, 517–530
  accountability and, 164, 525, 846–847
  algorithmic bias and, 132, 520
  child welfare determinations and, 522, 527, 844
  data collection and, 132, 141, 240, 527–528
  defined, 520–521
  disagreements on, 130, 134–136
  in facial recognition technology, 129, 522–523
  fintech and, 521–522
  in healthcare, 132, 212–214, 839, 844–846
  hiring practices and, 69–71, 131, 133, 139, 468, 521, 920
  incentive-based rating systems for, 518–519, 524–530
  mitigation strategies, 130, 136–141, 144–145n40, 572
  modeling and, 132–133, 139
  in predictive algorithms, 33, 129, 969
  proxy attributes and, 133, 141, 144n32
  in public AI systems, 428–429
  public benefits eligibility and, 522–523
  in public organizations, 392
  search engines and, 517–518
  task specification and, 131–132
  values-first approach to, 139–140
algorithmic decision-making
  accountability for, 168
  bias in, 33, 69–71, 129–133
  distributive outcomes of, 239–240
  in explainable AI, 173
  fairness in, 135, 138–139, 141–142
  flash crashes and, 49
  by lottery, 141–142
  in militaries, 919
  opacity of, 169, 969
  permit program review of, 304
  in public AI systems, 431–433
  thresholds for, 133, 136, 141, 142
  unintelligibility of, 164, 169
algorithmic impact assessments (AIAs), 173, 288, 500–501, 525, 528–529
Algorithmic Justice and Online Platform Transparency (AJOPT) Act (proposed), 572–573, 579
algorithms. *See also* algorithmic bias; algorithmic decision-making; predictive algorithms
  AlphaGo, 916, 930n2
  audits of, 264, 279, 290, 496, 500–503, 509–510, 528–529, 847
  for autonomous vehicles, 31, 32
  as black box technologies, 519–520, 522, 566
  COMPAS, 134, 136, 137, 280, 468
  content selection, 52, 54
  deep learning, 918, 923
  ethical issues and, 525
  fairness of, 33, 130, 136, 142, 352–353, 526
  in healthcare, 132, 212–214, 216, 839–853
  innovation in, 102n14, 842
  language models and, 203
  limitations of, 157
  microaggression detection, 238
  personalized, 560–561, 668
  in public AI systems, 424, 425, 428–429, 431–434
  recommender, 172, 203, 204, 206, 326–327, 332
  regulatory proposals, 172
  search, 203, 204, 206, 236, 843, 990
  transparency of, 481–491, 525, 526, 842, 846
alignment, 65–82. *See also* social alignment; value alignment
  complexity of, 31
  defined, 22, 65, 66, 323
  delegation and, 67–69, 68*f*
  direct, 65–66, 69–73, 77, 80–81
  harms of AI and, 682–683
  loyal AI systems and, 331, 332, 341n7, 341n10
  of NATO allies, 1016, 1021, 1022
  principal–agent, 30, 68–69, 68*f*
  system-related, 29, 30
ALJs (Administrative Law Judges), 781, 784, 788
Allegheny Family Screening Tool (AFST), 844
ALM (Autor-Levy-Murnane) hypothesis, 643–645, 647
Almeida, V., 399
AlphaGo algorithm, 916, 930n2
ALPRs (Automated License Plate Readers), 798
Alsinglawi, B., 540
Altman, S., 768

Amazon
Alexa, 195, 321, 327, 920
cashier-less grocery stores, 920
collective actions against, 588–589*t*, 592
employee monitoring by, 622, 675, 795n5
facial recognition technology, 508–509
in Global South, 989, 991
job candidate ranking, 50, 920
American Civil Liberties Union (ACLU), 502, 508, 510, 998
American Medical Association (AMA), 313, 342n12
Ananny, M., 486–487
anarchy, global, 34
Anderson, David, 260
Andini, M., 47–48
Anghie, Anthonie, 1001
Appeals Council Analysis Tool (ACAT), 783, 786, 788
Appiah, K. A., 228n7
Apple
algorithmic bias and, 143n8
Apple Watch, 840
business model, 335
consumer financial services, 860, 867–868
COVID-19 case monitoring and, 541
in Global South, 989
number of employees, 652, 713
Siri, 195, 321, 327, 920, 990
Arntz, M., 721n6
Arora, Payal, 995
Arrow, Kenneth J., 74
Arsenault, Amelia C., 17–18, 959
artificial intelligence (AI). *See also specific types of AI*
access to (*see* structured access)
as accountability instrument, 174–175
alignment of (*see* alignment)
assurance processes, 56, 1031
benefits of, 46–48, 56, 171, 421, 959
complexity of, 401, 620–631
definitions of, 39n1, 58n1, 101n9, 227n3, 254–255, 313n8, 341n1
economy and (*see* economic impact of AI)
ecosystem (*see* AI ecosystem)
existing and emerging applications, 95–98, 98*t*
explainable (*see* explainable AI)
field orientation, 734–742, 736*t*
fintech and, 521–522, 860–861, 867–871
forecasts on deployment of, 21
general, 115, 751, 916
in Global South, 987–993
governance of (*see* AI governance)
as GPT (*see* general purpose technologies)
harms of (*see* harms of AI)
in healthcare (*see* healthcare)
human extinction due to, 77, 99–100
implementation challenges, 96
as information technology, 26–29
as intelligence technology, 29–31
investment in, 35, 45, 95, 96, 102n13, 924, 959
limitations of, 96, 102n12
loyalty of (*see* loyal AI systems)
military applications (*see* militaries)
narrow, 621, 628, 915–916, 922
PAIS (*see* public AI systems)
politics and (*see* politics)
power exercised by, 202–207
progress monitoring, 57
public opinion on (*see* public opinion)
in public policy cycle, 537–545
race dynamics and, 35–38, 50, 310, 882–883, 891n6, 923, 925, 1038n56
risk clusters for, 22–24
safety issues (*see* safety)
SAI (*see* surveillance AI)
security issues (*see* security)
sociotechnical change and, 358, 361–362
structural injustice and, 210, 212–219
TEVV framework for, 56, 1031
transformative, 23, 103n20, 358
trustworthy, 412, 556–558, 563–564, 941–945, 947–952
Artificial Intelligence Act of 2021 (AIA)
on definition of AI, 255
documentation requirements, 266–267
on facial recognition technology, 825
on hazardous AI applications, 254
on high-risk AI systems, 265–266, 524, 577–578, 946–947, 954
horizontal AI regulation and, 937–938, 953
on proportionality, 266
on supply chain accountability, 266
on transparency obligations, 267

"White Paper" and, 261–262, 264–265, 945–946
Artificial Intelligence Decision-Making (AIDM). *See* automated decision-making
Artificial Intelligence Incident Database, 627
Asimov, Isaac, 739
Association for the Advancement of Artificial Intelligence (AAAI), 885
Atherosclerotic Cardiovascular Disease (ASCVD) Risk Estimator, 839
audits, 495–510
access to information for, 503–504
for accountability, 496, 497, 499–505, 500*f*, 510
actionable, 496, 502
algorithmic, 264, 279, 290, 496, 500–503, 509–510, 528–529, 847
case studies, 505–509
challenges of, 496–501, 622
for evaluation, 496–499
external, 495, 501–505, 502*t*, 507–510
internal, 495, 501–506, 502*t*, 510
of loyal AI systems, 334, 336, 339
of operational safety, 452
participatory, 452, 496, 497
of public AI systems, 428–429, 434, 435
standardization of, 504
visibility and impact of, 504–505
authoritarianism
digital, 34, 963, 967, 970
disenfranchisement and, 709
global trends, 968–970
information environment and, 717–719
military use of AI and, 966
regime strengthening, 97
surveillance AI and, 22, 34, 803, 826, 962, 969, 972
totalitarianism, 22, 28, 34, 89, 535
authority
colonialism and, 1001
for data processing, 868
for decision-making, 118, 125n11, 430
discretionary, 430–431
for exercise of power, 206, 207
moral, 159, 160
permit programs and, 299, 310
in public organizations, 384, 392, 394, 535
regulatory, 258, 259
automated decision-making, 276–291, 461–474
accountability and, 465–466
appeals process, 279–281
bias in, 464, 468, 517
control errors in, 447
deployment phase, 285–286
distributive outcomes of, 239–240
due diligence for, 284–286, 288, 289
explainability of, 866
factors in selection of, 470–473
human rights regarding, 277–284
inequality and, 468–469
justification of use, 462–464, 472–473
legality of, 461, 465–469, 474, 1029
litigation for harm caused by, 281–284, 290
loss of human control due to, 52
preconditions for use, 470–471
pre-deployment measures, 284–285
privacy and, 467–468
in public AI systems, 430–432
regulatory ecosystem for, 286–291, 287*f*
retirement of systems for, 286
shortcomings of, 897
transparency of, 466–467, 483
value of use, 471–472
Automated License Plate Readers (ALPRs), 798
automation. *See also* automated decision-making; robotic systems
bias and, 392, 464, 468, 517, 844
capital–labor complementarities and, 710
of data collection, 156
democracy and, 662, 681, 701–702, 715–717
economic impact, 53–54, 97, 99, 711–712, 714*t*, 964
in employment, 562–563, 644–647, 710–712
full, 99, 244, 711–713, 714*t*
in Global South, 721n6
harms of AI and, 670–674, 701–702
heavy, 242, 243, 245, 246
inequality and, 50–51, 54, 99, 468–469, 655, 728–729, 964, 970
language models and, 51
near-complete, 98–99, 103n21, 103n23, 711–712, 714*t*
public opinion on, 562–563
unemployment due to, 54, 102n16, 649, 921, 970
in workplace, 562–563, 644–647

autonomous vehicles
accident models for, 444–445
adaptability of, 185
algorithms for, 31, 32
benefits of, 171
ethics and, 32
implementation challenges, 96, 385
loyal AI systems and, 326
public opinion on, 554
safety issues, 31, 240, 444–445, 449
vulnerability to attacks or accidents, 49
autonomous weapons. *See also* lethal autonomous weapons
bans on, 55, 79, 171, 896, 898, 927
centralized control and, 28
drones, 49, 97, 354, 896, 918–920, 961, 965, 972
ethics and, 1025
harms of AI and, 682
human-in-the-loop capabilities, 35, 77, 961
in international law, 974
monitoring challenges, 178n34
totalitarianism and, 22
Autor, D. H., 102n16, 642, 643, 646–647, 649, 651, 652, 656n4, 670–672, 750, 754
Autor-Levy-Murnane (ALM) hypothesis, 643–645, 647
Ayres, Ian, 719, 722n19

B

Bajgar, Ondrej, 76
balance of power
diffusion of AI and, 922
digital authoritarianism and, 34
economic power and, 920
inclusivity proposals and, 997
in international politics, 914, 924–925, 959, 971–973
military technologies and, 707
personal data storage and, 152
Baldwin, James, 144n16
Balwit, Avital, 5–6, 65
bank–customer relationships, 861–863, 870
Bannister, F., 423, 426
Barbieri, D., 538–539
Bari, A., 540
Barocas, S., 144–145n40, 444
Bartels, Larry, 717
Bayesian additive regression trees (BART), 902–904, 905*f*, 910, 912n35
Bay of Pigs mission (1961), 462
Beck, Charlie, 832–833
behavioral manipulation, 661, 667–669, 691–693
Behavioural Learning for Adaptive Electronic Warfare (BLADE) program, 961
Bell, J., 299
Benjamens, S., 311
Benjamin, Ruha, 985
Benkler, Yochai, 829
Bennett Moses, Lyria, 362, 364–365, 373
Bergemann, Dirk, 664
Berners-Lee, Tim, 160
beta-testing, 986, 992, 995
Beyleveld, Deryck, 364
Bezos, Jeff, 202
bias
algorithmic (*see* algorithmic bias)
automation and, 392, 464, 468, 517, 844
availability, 463
cognitive, 178n38, 392, 462
confirmation, 463
in consumer finance, 869, 870
in decision-making, 30, 33, 69–71, 129–133, 462–463, 468, 682
disparate impact, 131–133, 144n27, 302, 523
disparate treatment, 131–133, 135, 218
in education, 130, 143n11
employment-related, 505–506
explicit, 517, 844
externalities and, 73
in facial recognition technology, 50, 129, 173, 507–509, 522–523, 558–559, 825, 969
gender (*see* gender bias)
in healthcare, 132, 212–214, 216, 217, 408, 839, 844–846, 852
hindsight, 448, 463
implicit, 213, 237–238, 428, 463, 844, 847
microaggressions, 227n2, 237–238
morally neutral, 143n2
optimism, 562
in political communities, 237
in public AI systems, 428–429, 431
quantification, 969

racial (*see* racial bias)
skill-biased technical change, 642–645, 729
smart city technologies and, 820–821, 823–826
systemic, 194, 625, 825, 971
Biber, E., 300, 311, 313n7
big brother effects, 662, 680–681, 699–701
big data
in COVID-19 pandemic, 540, 541
distributive outcomes of, 240
machine learning and, 186
for micro-targeting, 239
military use of, 920, 965
as policing tool, 830–833
predictive algorithms and, 144n24
privacy issues and, 159, 160
for RESAS, 389
smart city technologies and, 823–824, 827
societal implications of, 517
surveillance AI and, 800–801
transparency and, 483
bigger pie effect, 651–652, 656n5
Big Tech, 226, 572, 860, 867–868, 985, 989–990
Biometric Information Privacy Act of 2008 (BIPA), 307–308
Bird, Eleanor, 991
bisexuals. *See* LGBTQ communities
Bismark, Otto von, 463
black box technologies
accountability and, 173, 969
algorithms as, 519–520, 522, 566
evaluation challenges, 498
NIST reviews of, 314n20
permit programs for, 304
transparency and, 429, 467, 484, 486, 522, 573, 846, 969
Black Lives Matter (BLM) movement, 595, 828, 833
BLADE (Behavioural Learning for Adaptive Electronic Warfare) program, 961
blockchain, 174, 410
Boeing 737-MAX aircraft, 35, 454
Boix, Carles, 14, 707
Bostrom, Nick, 38, 83n10, 103n21, 751
Botero Arcila, Beatriz, 15, 820
bounded rationality, 383, 386, 390–391, 394
Bovens, Mark, 390
Bradford, Anu, 258, 259
brain drain, 741
Braithwaite, J., 576, 579
Brand, J. E., 769
Brauneis, R., 285
Brayne, Sara, 826
Brazil, regulatory framework for AI in, 578
Bresnahan, T. F., 87
Brexit referendum (2016), 865, 871, 962
Brockman, John, 394
Brooke, S. J., 143n12
Brookman, D. E., 740
Brown, Chad, 888
Brown, Joshua, 445
Brownsword, Roger, 202, 361–362, 364
Bruney, Chris, 283
Brussels Effect, 254, 258–259, 268
Brykman, Steven, 278
Brynjolfsson, Erik, 96–98
Bryson, Joanna J., 8, 253, 255
Buiten, M. C., 402
Bullock, Justin B., 1, 9–10, 383, 390, 400, 421
Buolamwini, J., 507
Burrell, J., 483–484
Buzan, Barry, 889, 890

C

CAFC (Court of Appeals for the Federal Circuit), 308, 315–316n39
Caines, Stephen, 15, 797
Calcaño, Alejandro, 1001
calibration, 33, 137, 141, 145n54, 353, 454, 843
Cambridge Analytica, 102n11, 154, 316n42, 962, 992
Campaign to Stop Killer Robots, 7, 561, 897, 904, 910
Canfil, Justin Key, 17, 895, 924
capital–labor complementarities, 709–713
Capurro, Rafael, 987
Carlsmith, J., 751
CCW (Convention on Certain Conventional Weapons), 895–909, 926, 1016
Challenger explosion (1986), 462
Chan, Alan, 993, 1002
changing pie effect, 652, 654

chatbots
DialoGPT, 609
identification requirement for, 278
natural language processing for, 422
non-state actors and, 972–973
OpenAI GPT- 3 and, 53
public agencies and, 430, 473
as therapists, 326
in vaccine promotion, 51
Chen, Frank, 738, 742
Chen, Yu-Che, 1, 9, 10, 421
child welfare determinations, 522, 527, 844
Chimni, B. S., 1002
China. *See also* U.S.–China relations
Belt and Road Initiative, 813*f*, 983
at CCW meetings, 897
collective actions in tech industry, 594–595
COVID-19 pandemic in, 540, 542
decoupling from Western AI, 35, 881
facial recognition technology in, 51, 58n4, 559, 803, 962
Golden Shield Project, 718–719
healthcare regulation, 847, 848
human rights violations in, 891n12
leadership in AI, 959, 967, 972, 983
nuclear arsenal, 38
regulation of AI in, 268
trade practices, 881, 889
Uighurs in, 51, 58n4, 803
Choi, J. P., 664
Christian, Brian, 72
Christiano, Paul, 38, 71–72
Churchman, C. West, 535
CIR (Common Identity Repository), 800–801
Cisco, 822
citizen suits, 307
Citron, Danielle, 199
Clark, Jack, 9, 345
Clarke, Arthur C., 772
Clarke, Roger, 199
Clarke, Sam, 5, 45
classifiers. *See* algorithms
Clean Air Act Amendments of 1990, 314n16, 316n50, 525
Clemens, M. A., 733
climate change, 34, 45, 47, 56, 288, 566, 596
Clinton, Bill, 648
cloud-based systems
EU infrastructure for, 949, 952
safety considerations, 451
structured access and, 605, 608, 610, 612–616, 616n1
surveillance AI and, 799, 801
transparency and, 483
Coffee, M., 540
Cofone, I. N., 485
Coglianese, Cary, 11, 461, 488
Cognitec, external audit of, 509
Cognitive Trade Advisor (CTA), 963
Collingridge, David, 313n3, 1018, 1036n23
Collingridge dilemma, 299
colonialism, 210, 223, 838, 983, 987–988, 999–1003
command-and-control governance, 574
Common Identity Repository (CIR), 800–801
communication. *See also* information and communication technologies
digitally mediated, 22
formal and informal, 383, 390, 392–394
harms of AI and, 662, 679–680, 698–699
horizontal, 203, 204
human–machine, 26
political, 89, 679–680
in public organizations, 383, 386, 390, 392–395
question exploration tool on, 117
strategic, 962, 975, 1027
Community Oriented Policing Services (COPS), 831–832
comparative advantage, 68, 646, 654, 674, 694, 711, 733, 747, 756, 921, 1032
COMPAS algorithm, 134, 136, 137, 280, 468
competition
AI race, 35–38, 50, 310, 882–883, 891n6, 923, 925, 1038n56
first-mover advantage, 922, 929, 961, 972, 1001, 1002, 1027
FTC regulation of, 309
by great powers, 23, 34–39, 897, 922
market, 27, 38, 261, 707
perfect, 773n6
in public policy cycle, 536
revealed, 241
social norms and, 76
technological, 564, 881, 888, 1021
unfair, 661, 666–667, 689–691
U.S.–China, 881–883, 886

value erosion from, 23, 38–39
complementing force, 642, 646–648, 650–653
computers. *See also* computing power
adoption of, 89–90, 100–101nn2–3
ALM hypothesis and, 644
as general purpose technology, 25, 86, 87, 89–90
military use of, 38, 89, 90
social significance of, 387–388
supercomputers, 400, 751, 891n6, 950
vision systems, 49, 345–349, 388, 807, 886, 918, 992
computing power
access to, 56
for automated decision-making, 470
bounded rationality and, 386
concentration of, 50
of human brain, 21
machine learning and, 198, 868
for predictive capabilities, 972
as technical bottleneck, 96, 102n14
Condorcet, Marquis de, 74
confidentiality. *See* privacy
conflict. *See also* warfare
avoidance of, 152
domestic, 33–34, 47
political, 32, 33, 970
severity of, 48–49
social, 25, 32, 33, 709–711
threshold for, 974
conflicts of interest, 241, 320–328, 335–336, 339–340, 505, 802
Connolly, R., 423, 426
constraint-based accident models, 444–445
constructivism, 486–487, 489–491, 1020
consumer financial privacy, 860–872
bank confidentiality and origins of, 861–863, 870
data protection regulation and evolution of, 863–867, 870–871
fintech and, 860–861, 867–871
future research needs, 871–872
Consumer Financial Protection Bureau, 314n21
contact tracing, 330, 798, 808
control problem, 22, 29–31, 69, 489
conveners, 110–114, 124. *See also* multi-stakeholder convening
Convention on Certain Conventional Weapons (CCW), 895–909, 926, 1016
Conway's law, 741
cooperation. *See also* alignment
barriers to, 52
economic, 243
informal social norms and, 78
international, 48, 578, 628, 928, 963, 974, 984
of NATO allies, 1016, 1017, 1021, 1022, 1024, 1029
in political communities, 234
social, 226, 228n18, 235, 237–239, 242–243, 245
COPS (Community Oriented Policing Services), 831–832
copyright law, 32, 500, 619, 773
Costa, Maren, 599n10
Costeja, Mario, 157
Court of Appeals for the Federal Circuit (CAFC), 308, 315–316n39
COVID-19 pandemic
contact tracing and, 330
detection and monitoring of cases, 540, 541
governmental failures in, 462
information sources during, 203
isolation of cases, 540, 542
mask hesitancy during, 52
public policy cycle during, 535, 536, 539–545
social distancing and hygiene measures, 541, 542
surveillance AI during, 809
transmission predictions, 47, 542
unemployment during, 522, 649
vaccines during, 51, 52, 541–544, 990
Crawford, Kate, 486–487, 991
credit scores and reports, 131, 140, 187, 519, 521, 862, 868–869
criminal justice system. *See also* policing
algorithm use in, 33, 129, 134, 528
power of AI in, 202
racial bias in, 463
risk assessment tools in, 129, 140
Cross, Steve, 738
cryptocurrency, 174–175
CTA (Cognitive Trade Advisor), 963
Cunningham, Emily, 599n10
curse of flexibility, 450

Cusicanqui, Sylvia, 985
cybersecurity
for automated decision-making, 470
in China, 884, 887
data breaches as failures of, 158
designing systems for, 333
EU funding for, 950
hybrid governance models and, 576
impact assessments on, 56
information-sharing platforms, 389
machine learning and, 971
permit programs and, 312
public AI systems and, 426
structured access and, 607
surveillance AI and, 811*f*, 813
threat analysis for, 422
cyberwarfare, 35–36, 965, 974

D

Dafoe, Allan, 5, 21, 24, 37–38, 48, 562, 1019, 1020
Danaher, John, 175
Daniels, N., 54
Danks, David, 7, 183
Danzig, R., 29
Dash, J. Michael, 983
data brokers, 141, 151, 153, 157–159, 801
data capitalism, 242–243, 245, 983, 987, 1001
data collection
algorithmic bias and, 132, 141, 240, 527–528
automation of, 156
conflict escalation and, 49
consent to, 159–160
decision-making on, 117–118
in Global South, 990–991
incentives for, 51, 141, 157
minimization of, 155–156, 865
notice requirements on, 119, 122–123
opting out of, 122–123
permit programs for, 302
privacy issues in, 528
standardization of, 846
storage limitations, 156–157
submodularity and, 664–666
for surveillance AI, 800
transparency of, 122, 572
unavoidability of, 123, 125n14
data economy, 141, 149, 152, 242, 246, 989–991, 1002
datafication, 242–243, 245, 246, 867, 994
data mining, 276, 316n42, 537, 802, 918
Data Protection Acts (DPAs), 864–868
Data Protection Directive of 1995, 865, 866
data protection impact assessments (DPIAs), 865
data quality, 132, 141, 407, 411, 849
decision-making. *See also* algorithmic decision-making; automated decision-making
accountability for, 54, 58, 168
accuracy of, 140, 464, 485
authority for, 118, 125n11, 430
bias in, 30, 33, 69–71, 129–133, 462–463, 468, 682
collective, 199, 462–463, 575
on data collection, 117–118
decoupling processes for, 140
democratic, 235, 236, 241, 802
ethical, 54, 74
fairness in, 135, 138–139, 141–142, 444
in healthcare, 838, 839, 846
limits on AI and, 117–119
loyal AI systems and, 332
machine learning as tool for, 131, 384
in militaries, 897, 919, 965–966
in networked governance, 575
in political communities, 234, 235
power and, 236, 982, 985, 993, 995, 997
procedures for, 54, 57–58, 59n8
in public organizations, 384–386, 390–395
on public services, 430–432
shortcomings of, 462–463
smart city technologies and, 821, 824
social choice theory on, 58–59n7
by technical experts, 111
time-constrained, 390
transparency of, 58, 466–467, 482, 483, 486–487
deepfake technology, 352, 353*f*, 357n10, 360, 962, 968, 971–972, 1030
deep learning (DL)
algorithms and, 918, 923
applications of, 102n14
diagnostic tools and, 839–840, 887

in explainable AI, 173
fintech and, 868, 869
public policy cycle and, 537
structured access and, 610, 611, 611*t*
DeepMind, 46, 47, 52, 612, 736*t*, 773, 773n2, 930
Defense Department, U.S. (DoD)
on AI development, 50, 1031, 1041n86
Joint All-Domain Command and Control, 384
on lethal autonomous weapons, 34–35
Project Maven and, 593, 918, 961, 965
DEI (diversity, equity, and inclusion), 211–212, 219–224, 227n4. *See also* equality; inclusivity
Dekker, Sidney, 453–454
delegation
alignment and, 67–69, 68*f*
examples of, 68, 82n8
of power, 165, 166, 168, 465
in public organizations, 385
democracy. *See also* elections
accountability and, 167, 175, 176, 998–999
automation and, 662, 681, 701–702, 715–717
capitalism and, 710, 716
early vs. modern, 101n8
economic impact of AI and, 715–717
equality and, 22, 152
foundations of, 708–709
global trends, 92, 93*f*, 715–717, 968–971
harms of AI and, 662, 676, 681, 682, 701–702
income levels and, 715–716
industrialization and, 94
justice and, 152
liberal, 152, 205–206, 239, 480, 960, 963–964, 967–971
military use of AI and, 965–966
mitigation of AI's negative effects on, 719–720
near-complete automation and, 99, 103n23
privacy rights and, 152, 160
re-equilibration process and, 716–717
representative, 73, 707
technocracy vs., 682
Third Wave of, 676
Dempsey, Gaia, 8–9, 320
Dempsey, Mark, 8, 253
dependency theory, 984–985
developed countries. *See* Global North
developing countries. *See* Global South
DHS (Homeland Security Department, U.S.), 831–833
Diakopoulos, Nicholas, 496
Diamond, Larry, 718
Diao, X., 732–733
digital divide, 364, 424, 538, 990
Digital Innovation Hubs (DIHs), 944, 948, 950, 951
Digital Services Act of 2020 (DSA), 254, 260–264, 504
Digital Single Market (DSM), 257
digital sovereignty, 948–951, 988–989, 995, 1003
Dinerstein v. Google (2023), 851
Ding, Jeffrey, 17, 881
direct alignment, 65–66, 69–73, 77, 80–81
disaster response, 47, 964
discrimination. *See* bias
disease. *See also* COVID-19 pandemic; healthcare
drug design for, 46, 97, 773n2
outbreak warning systems, 47
risk prediction capabilities, 46, 839
vaccines against, 51–52, 100n1, 541–544, 554, 808, 990
disinformation
algorithms and, 407
deepfakes and, 962, 968, 971
fake news, 33, 239, 407, 411, 541–542, 610, 677, 719
health-related, 52
mass production of, 605
as soft power, 967
targeted, 960, 962, 968
as threat to democracy, 968–969
disloyalty, 322–325, 328, 329, 335, 340
diversity, equity, and inclusion (DEI), 211–212, 219–224, 227n4. *See also* equality; inclusivity
divided loyalty, 322–325, 330, 340
DL. *See* deep learning
Dobbe, Roel I. J., 9, 10, 441, 442, 445
DoD. *See* Defense Department, U.S.

DOE (Energy Department, U.S.), 518, 525
Dorn, David, 652
DPAs (Data Protection Acts), 864–868
DPIAs (data protection impact assessments), 865
Drexler, K. E., 103n21, 617n3
driverless cars. *See* autonomous vehicles
drones, 49, 97, 354, 539–543, 896, 918–920, 961, 965, 972
DSA (Digital Services Act of 2020), 254, 260–264, 504
DSM (Digital Single Market), 257
due process
  algorithms and, 283
  power and, 199, 206
  procedural, 466
  public AI systems and, 431
  as safety constraint, 447
  in SSA Disability Program, 781, 782
duty of care, 173–174, 519, 872
Dwivedi, Y. K., 399

E

EC. *See* European Commission
echo chambers, 52, 73, 241, 244, 407, 411, 662, 676–679, 968, 971
ECHR (European Convention on Human Rights), 260, 447, 863
eCMS (electronic case management system), 782–783, 788, 791–792
economic impact of AI
  asset ownership and, 713–714
  automation and, 53–54, 97, 99, 711–712, 714*t*, 964
  capital–labor complementarities and, 709–713
  existing and emerging applications, 95, 97, 98*t*
  flash crashes and, 29, 49
  political consequences, 714–717
  productivity, 25, 87, 88*f*, 91, 920–921
  relative weight of capital and, 712–713
  skepticism regarding, 103n17
  supply chains and, 102n10
economic inequality, 103n16, 729, 823
economic power
  general purpose technologies and, 916
  Global South and, 994, 1002
  military power and, 916–917, 920
  personal data trade and, 158
  structural injustice and, 211, 217, 220
economy. *See also* economic impact of AI
  complete markets, 764–765
  data, 141, 149, 152, 242, 246, 989–991, 1002
  equity markets, 774n8
  general purpose technologies and, 25, 87–89, 88*f*, 94
  global, 732, 889
  information technologies and, 26–27
  political, 829–833, 999–1000
  revolutionary technologies and, 87, 90, 94
  total factor productivity and, 25, 87, 88*f*, 91
Edgerton, David, 882
EDTs (Emerging and Disruptive Technologies), 1015–1016, 1021–1022, 1027–1033
education
  bias in, 130, 143n11
  on healthcare AI, 841–842, 852
  income level and, 729
  inequality of opportunities for, 210
  personalized, 46
  skills mismatch and, 648
  trust in AI and level of, 557
EESC (European Economic and Social Committee), 938–940, 943
EGE (European Group on Ethics in Science and New Technologies), 938, 939
E-Government Act of 2002, 467
EIAs (environmental impact assessments), 77–78, 495
elections. *See also* democracy
  legitimacy of, 153
  in modern democracy, 101n8
  security of voting systems in, 174
  U.S. presidential election (2016), 102n11, 154, 562, 962
  voter behavior in, 102n11
electricity, 25, 47, 86–87, 89, 94, 98, 100n1, 101n3
electronic case management system (eCMS), 782–783, 788, 791–792
ELISE (European Network of AI Excellence Centres), 952

ELLIS (European Laboratory for Learning and Intelligent Systems), 952
EMA (European Medicines Agency), 847, 849–850
Emerging and Disruptive Technologies (EDTs), 1015–1016, 1021–1022, 1027–1033
Emerson, K., 423
Emotion AI. *See* affective computing
employment. *See also* hiring practices; unemployment
  AI ethicists, 586–587, 592–594, 597
  AI workers, 584–587, 592–597, 595*t*
  allocation of, 756–757, 759–761, 759*f*, 764
  ALM hypothesis and, 643–645, 647
  amenities from, 761–762, 762*f*, 767
  automation in, 562–563, 644–647, 710–712
  blue-collar jobs, 669
  capital–labor complementarities, 709–713
  collective actions in tech industry, 587–597, 588–591*t*, 598n5, 599n10
  discrimination in, 505–506
  economic redundancy of labor, 751–753, 768–771
  externalities and internalities, 763–764
  full automation in, 99, 244, 711–713, 714*t*
  globalization of, 883–884, 884*t*, 886
  harms of AI and, 661, 669–676, 694–695
  history of technology and, 642–648
  for individual agents, 757–759, 757*f*
  job polarization in, 102–103n16
  labor-saving AI, 727, 730, 732–733
  machines as perfect substitute for labor, 750–751
  monitoring of employees, 622, 675–676, 730–731, 795n5
  near-complete automation and, 98–99, 711–712, 714*t*
  nostalgic jobs, 755
  pink-collar jobs, 649
  skill-biased technical change in, 642–645, 729
  surveillance AI and, 202, 220, 622
  task creation and, 653–655
  technological progress and, 749–750, 749*f*
  white-collar jobs, 384, 558, 649, 671, 710
  winner-takes-most labor markets, 22, 27
  for women, 89, 224, 527
energy capture, 90–92, 91*f*, 99, 101nn6–7
Energy Department, U.S. (DOE), 518, 525
Energy Star Rating program, 518–519, 525
Engler, Alex, 277, 528
environmental degradation, 991–992
environmental impact assessments (EIAs), 77–78, 495
Environmental Protection Agency, U.S. (EPA), 301, 310, 315n25, 518, 525, 526
EP. *See* European Parliament
equality. *See also* inequality
  of consumption, 27
  democracy and, 22, 152
  erosion of, 38, 152
  of healthcare, 838, 840, 845, 847, 852
  political, 175, 236, 243, 246
  proportional, 135
  public AI systems and, 426, 428, 431, 432, 435
  social, 205–206, 236, 243, 246
  tradeoffs involving, 384
Equifax data breach (2017), 158
equity. *See* diversity, equity, and inclusion
ethics. *See also* value alignment
  accountability and, 165, 169–170, 173–174, 176, 221
  affective computing and, 109–115, 124
  AI ethicists, 586–587, 592–594, 597
  algorithms and, 525
  autonomous vehicles and, 32
  autonomous weapons and, 1025
  benchmarks related to, 350–351, 351*f*
  codes of, 78, 212, 342n12, 938
  decision-making and, 54, 74
  duty of care, 173–174, 519, 872
  in EU governance of AI, 926, 938–943, 945, 946
  fiduciary duties, 325–326, 335, 337–340, 342nn13–15
  in healthcare, 54, 843
  moral agency, 82n4, 152
  moral progress, 48
  in NATO governance of AI, 1025–1028, 1026*t*
  politics and, 222
  question exploration tool for, 114–117
  surveillance AI and, 797–798, 814, 815

EU. *See* European Union
Eubanks, Virginia, 239, 240
European Commission (EC)
advisory bodies to, 938
AIA (*see* Artificial Intelligence Act of 2021)
AI HLEG, 848, 940–943, 945
"Artificial Intelligence for Europe," 264, 939
"Building Trust in Human-Centric AI," 943–944
"Communication on AI for Europe," 939, 940, 943, 949
on conformity assessments, 288
"Coordinated Plan on AI," 939–941, 944, 945, 948–950, 952
"Digital Day Declaration on Cooperation on AI," 939–940, 955n5
Digital Services Act, 254, 260–264, 504
Digital Single Market strategy, 257
e-Commerce Directive, 261, 262
"European Approach to Artificial Intelligence," 261
"European Strategy for Data," 944, 949
general permission regulatory approach, 311, 312*t*
on human oversight, 448
stakeholder consultation process, 955n9
von der Leyen as president of, 253–254, 257, 944
"White Paper on Artificial Intelligence," 261–262, 264–265, 848, 849, 944–952
European Convention on Human Rights (ECHR), 260, 447, 863
European Court of Justice, 827
European Economic and Social Committee (EESC), 938–940, 943
European Group on Ethics in Science and New Technologies (EGE), 938, 939
European Laboratory for Learning and Intelligent Systems (ELLIS), 952
European Medicines Agency (EMA), 847, 849–850
European Network of AI Excellence Centres (ELISE), 952
European Parliament (EP)
Committee on Industry, Research and Energy, 941
"A Comprehensive European Industrial Policy on Artificial Intelligence and Robotics," 953
ethical framework for AI adopted by, 945
"Recommendations to the Commission on Civil Law Rules on Robotics," 265, 938–940, 943, 952–953
European Union (EU). *See also* General Data Protection Regulation
AI Lighthouse Centres in, 952
AI megaprojects in, 951–952
AI observatory proposal, 953–954
Brussels Effect and, 254, 258–259, 268
Charter of Fundamental Rights, 263, 940, 942
Data Protection Directive, 865, 866
Digital Innovation Hubs, 944, 948, 950, 951
digital sovereignty in, 948–951
ethical approach to AI, 926, 938–943, 945, 946
foreign direct investment framework, 951
formative AI policy developments, 938–941
future of AI governance in, 951–954
healthcare regulation, 847–850
human-centric approach to AI, 938–939, 941, 943–944, 948, 951
on privacy as fundamental right, 260–261
regulatory capacity, 258–259, 269n17
right to be forgotten in, 157, 866, 923
risk-based framework for AI regulation, 337
Single Market, 257, 258, 269n12, 269n17
standardization efforts, 954
transnational AI governance by, 253–254, 257–268
on trustworthy AI, 412, 941–945, 947–952
Evans, Benedict, 256
event-based accident models, 444
explainable AI (XAI), 183–195
accountability and, 172–173
algorithmic decision-making in, 173
consumer finance and, 869–871
explanation-generating, 187, 190, 193, 194
General Data Protection Regulation on, 928
governance and, 183–195
human-explainable, 187–188, 193, 194

human-interpretable, 188, 190, 193–195
public AI systems and, 430, 431
taxonomy of, 186–188
theories of, 185–186, 188–192
transparency and, 481, 488
externalities
data, 664–666, 684
employment, 763–764
negative, 23, 31, 76, 629, 666, 733, 763, 769
pecuniary, 73, 696
positive, 31, 73, 748, 763, 769
risks associated with, 172
social alignment and, 65, 69–81
external validity, 144n36, 471, 839

F

Facebook
algorithmic bias and, 518, 523–524
Cambridge Analytica scandal and, 102n11, 154, 316n42
collective actions against, 589*t*, 591*t*, 592
congressional hearings on, 38
content selection algorithms, 54
data mining by, 316n42
facial recognition technology, 509
friend-suggestion tool, 159
in Global South, 989
interest categories assigned by, 560
internal audits by, 495
model drift and, 843
number of employees, 713
personalized algorithms, 668
recommender systems, 95, 204
workforce diversity, 527
facial recognition technology (FRT)
accuracy of, 240, 385, 507, 508, 825
bans on, 171, 508, 509, 574, 604, 825, 998
bias in, 50, 129, 173, 507–509, 522–523, 558–559, 825, 969
data collection for, 800
deceptive data use involving, 307, 310
Gender Shades study on, 507–509
hiring practices and, 506, 559
on lethal autonomous weapons, 961
in policing, 51, 95, 202, 469, 507–509, 523, 559, 799
public opinion on, 554, 558–559
structured access, 605
as surveillance AI, 51, 58n4, 509, 797, 803, 962, 969
Facial Recognition Vendor Test (FRVT), 355, 508
fairness
algorithmic, 33, 130, 136, 142, 352–353, 526
in decision-making, 135, 138–139, 141–142, 444
in EU's trustworthy AI, 942, 943
in healthcare, 838, 842, 844, 846
of hiring practices, 506
limitations of, 218–219
of permit programs, 314n9
political communities and, 237, 239
of predictive algorithms, 136, 442
public AI systems and, 426, 428, 431, 435
rights-based approach to, 519
social norms and, 80
as standard of justice, 134, 135, 138
statistical, 136–137, 145n48
fake news, 33, 239, 407, 411, 541–542, 610, 677, 719
false negatives, 33, 136–137, 145n50
false positives, 33, 136–137, 145n50, 392, 691–693, 809, 849, 902
Fan, W., 841
FDA. *See* Food and Drug Administration, U.S.
FDI (foreign direct investment), 951
Federal Bureau of Investigation v. Fazaga (2022), 806
federalism, 309, 310
Federal Trade Commission, U.S. (FTC), 307, 309–310, 316nn41–42, 342n19, 518, 521, 525, 578–579, 814
females. *See* women
fiduciary duties, 325–326, 335, 337–340, 342nn13–15
filter bubbles, 33, 52, 239, 241, 244, 407, 411, 560, 677, 685
financial privacy. *See* consumer financial privacy
Fino, Steven, 1020
fintech, 521–522, 860–861, 867–871
fishbowl transparency, 466–467, 481
Flesch-Kincaid readability scores, 906

Floridi, Luciano, 166, 399
Food and Drug Administration, U.S. (FDA)
  AI permit reviews by, 314n21
  automated decision-making and, 464
  external audit function of, 502
  medical device regulation, 282, 300, 302, 311, 314n17, 847–850
  nutrition labels from, 524
  wearable device approval, 840
foreign direct investment (FDI), 951
Foucault, M., 203
Fourcade, M., 144n24
Fourth Industrial Revolution (4IR), 988, 1002–1003
Framingham 10-year Risk Score, 839
Freedom of Information Act, 467, 479
Freire, Paolo, 985
Frey, Carl Benedikt, 711
frictional technological unemployment, 648–649
Friedman, Milton, 770
FRT. *See* facial recognition technology
FRVT (Facial Recognition Vendor Test), 355, 508
Frye, M., 144n18
FTC. *See* Federal Trade Commission, U.S.
full automation, 99, 244, 711–713, 714*t*
function creep, 799, 801, 808, 809, 815
Fung, A., 526

G

Gabriel, Iason, 226, 333
Gannon, Andres, 899
Garcia, Eugenio Vargas, 984
Garfinkel, Ben, 6, 86
Gartenstein-Ross, D., 972
Gasser, U., 399
gays. *See* LGBTQ communities
GDPR. *See* General Data Protection Regulation
Gebru, Timnit, 507, 593
Gellert, R., 482
Gelman, Andrew, 902
gender bias. *See also* women
  credit limits and, 143n8
  decision-making impacted by, 463
  in facial recognition technology, 522–523, 558
  in global data sets, 844
  hiring algorithms and, 468, 920
  knowledge sharing and, 143n12
  in policing, 408
  reinforcement of, 144n18
  search engines and, 517
  structural injustice and, 212, 220, 224
gender justice, 135, 141
Gender Shades study, 507–509
General Data Protection Regulation (GDPR)
  on anonymized data, 159, 850–851
  on automated decision-making, 279
  on data minimization, 155
  ethical issues and, 927
  on explainable AI, 928
  on health-related data, 850
  national security cutouts, 928
  on privacy, 254, 260–261, 332, 411, 865–868, 923
  on pseudonymized data, 851
  on right to be forgotten, 75, 157, 866, 923
  on storage of personal data, 156
general purpose technologies (GPTs)
  definitions of, 24–25, 916
  economic impact, 25, 87–89, 88*f*, 94
  historical examples, 25, 86, 87, 88*t*
  human trajectory shifts and, 90–94, 91*f*, 93*f*
  innovation and, 24–26, 86, 87, 921
  national power and, 25, 914
  origins of, 89–90, 96
  proliferation rates for, 972
  properties of, 21, 25–26, 86, 87
  revolutionary, 25, 87, 90–94, 94*b*, 98
General Services Administration, U.S. (GSA), 384
George, Henry, 771–772
Gereffi, G., 574
GGE (Group of Governmental Experts), 896–899, 911, 926
Gigerenzer, Gerd, 625, 630
GitHub, 53, 347, 425, 605, 610, 616
Glaze, Kurt, 15, 779
Glissant, Édouard, 983

global financial crisis (2007–2009), 255–256, 830, 832, 867
Global Indigenous Data Alliance, 999
Global North
  AI systems in, 989–990
  democracy in, 715
  dependency theory and, 985
  power structures in, 1000
  structural injustice in, 210
  supply chain ignorance in, 996
  surveillance capitalism and, 990
  terminology considerations, 982–983
Global Partnership on AI (GPAI), 35, 256, 939, 984, 991, 1034–1035n10
Global South, 981–1003
  agency of, 986
  AI systems in, 987–993
  automation in, 721n6
  coalition building and solidarity in, 999
  coloniality of power in, 1000–1002
  commercial exploitation in, 992, 995, 1003
  democracy in, 717
  dependency theory and, 984–985
  digital sovereignty in, 988–989, 995, 1003
  extractive logics in, 990–992
  inequality in, 717, 982, 985–986, 994, 997, 1002
  intellectual property in, 990
  labor markets and workers in, 992
  paradox of participation and, 985–986, 1003
  as peripheral to governance, 986–987
  political economy of resistance in, 999–1000
  role in global AI governance, 927, 993–1000
  smart city technologies and, 822
  terminology considerations, 982–983
global warming. *See* climate change
Godfrey, G., 870
Goodfellow, T., 575–576
Goodman, E. P., 285
Goodwin, J., 307
Google
  AI assistant, 327
  algorithmic bias and, 352–353, 517, 518
  collective actions against, 588–591*t*, 592, 599n8
  COVID-19 case monitoring and, 541
  ethical commitments from, 412
  Flu Trends, 843
  in Global South, 989, 991
  image recognition tool, 845
  internal audits by, 495, 502
  learning algorithms, 930n4
  licensing agreements, 885
  number of employees, 713
  standards alliances, 884
  tensor processing unit, 868
  workforce diversity, 527
Google AdSense, 132
Google Cloud, 605
Google Health, 46
Google Search, 157, 556, 610, 739
Google Translate, 889
Gordon, Robert, 103n17
governance. *See* AI governance
GPAI (Global Partnership on AI), 35, 256, 939, 984, 991, 1034–1035n10
GPTs. *See* general purpose technologies
Grace, K., 751
Graetz, Georg, 921
Grant, Ruth W., 167
great powers
  competition among, 23, 34–39, 897, 922
  military AI debates, 907*f*, 908, 908*f*, 911
  strategic general purpose technologies of, 26
  warfare among, 22, 924–925
Green, Ben, 448, 828, 829, 984
Greene, K. Gretchen, 6, 109
Grissom, Adam, 1019
Group of Governmental Experts (GGE), 896–899, 911, 926
groupthink, 463
GSA (General Services Administration, U.S.), 384
Gurses, S., 451
Guyger, Amber, 216
Guzmán, Rigoberto Lara, 982

H

Haataja, Meeri, 8, 253
Hampson, Roger, 158

Hanson, Jon D., 668
Hanson, Robin, 38, 101n4
harms of AI, 48–53, 660–702. *See also* bias
  alignment problem, 682–683
  autonomous weapons and, 682
  behavioral manipulation, 661, 667–669, 691–693
  big brother effects, 662, 680–681, 699–701
  collective, 154–155, 160, 220
  communication and, 662, 679–680, 698–699
  complexity of, 622–624
  conflict severity increases, 48–49
  contextual incompatibility, 992–993
  control of information and, 661, 663–669
  democracy and, 662, 676, 681, 682, 701–702
  environmental degradation, 991–992
  excessive automation and, 670–674, 701–702
  historic assessment of, 996
  homogeneity and, 123–124, 126nn15–16
  human judgment and, 674–675, 696
  labor market effects, 661, 669–676, 694–695
  legitimate interpretations of, 995–996
  loss of human control, 52–53
  mitigation efforts, 55–56, 683–686, 813–814
  political discourse and, 676–681
  power concentration and, 50–51, 54
  privacy violations, 73, 79, 150–155, 661, 664–666, 687–689
  problem-solving ability, undermining of, 51–52
  reporting process for, 585, 587, 592–594, 592*t*, 597–598
  socioeconomic, 623–624
  unfair competition, 661, 666–667, 689–691
  vulnerability to attacks or accidents, 49–50, 622
  whole-system view of, 996
Harrison, T. M., 485–488
Hauge, Janice Alane, 989
Hawking, Stephen, 398
hazard analysis, 441, 443–444, 443*f*, 450–451
Heads of Medicines Agencies, 847
Heald, D., 480–481
healthcare, 838–853. *See also* disease
  algorithms in, 132, 212–214, 216, 839–853
  bias in, 132, 212–214, 216, 217, 408, 839, 844–846, 852
  chatbot therapists, 326
  clinical AI applications, 839–840
  decision-making in, 838, 839, 846
  diagnostic tools, 171, 178n33, 839–840, 887
  equality of, 838, 840, 845, 847, 852
  ethical issues, 54, 843
  fairness in, 838, 842, 844, 846
  inclusivity of, 838, 840, 845, 847, 852
  innovation in, 838, 842, 844, 852
  lifecycle of AI in, 840–844
  medical devices, 282, 300, 302, 311–313, 314n17, 847–850
  precision medicine, 46, 330, 891n6
  privacy issues, 840, 843, 850–851
  Quintuple Aim of, 838, 840, 844, 852
  recommendations and hypotheses, 851–853
  regulatory frameworks, 847–851
  safety in, 841, 842, 847, 848, 852–853
  structural injustice in, 212–214, 217
  telemedicine, 540–542
Health Insurance Portability and Accountability Act of 1996 (HIPAA), 630, 851
Healy, K., 144n24
heat lists, 823, 824, 829
Heaven, W. D., 543
Helleiner, Eric, 986
Hellman, D., 131
Hicks, J., 749
Higgins, Brian Wm., 8, 299
High-Level Expert Group on Artificial Intelligence (AI HLEG), 848, 940–943, 945
Himmelreich, Johannes, 1, 7, 210
HIPAA (Health Insurance Portability and Accountability Act of 1996), 630, 851
HireVue, 506, 521
hiring practices
  affirmative action and, 139, 218
  algorithmic bias in, 69–71, 131, 133, 139, 468, 521, 920
  Emotion AI and, 521
  facial recognition technology and, 506, 559
  incentive-based rating systems for, 527
  internal audits of, 505–506
  privacy violations and, 153–154
Ho, Daniel E., 15, 779
Holmes, O. W., 309

Homeland Security Department, U.S. (DHS), 831–833
homosexuality. *See* LGBTQ communities
Hood, C., 480
Horenovsky, Jan, 76
Horizon Europe, 948, 950, 955n18
Horowitz, Michael C., 17, 914, 970–972
horse equilibrium, 653–654
Housing and Urban Development Department, U.S. (HUD), 523, 573
Hu, L., 444
Huang, Hsini, 9–10, 383
hub and spoke networks, 113, 113*f*
Hubinger, Evan, 72
Hudson, Valerie M., 1, 8, 276
Hughes, Thomas, 1020
human-explainable AI, 187–188, 193, 194
human-interpretable AI, 188, 190, 193–195
humanitarian law, 326, 342n18, 897, 919, 927, 974, 1027, 1029
human rights
  affective computing and, 109
  algorithmic bias and, 520
  automated decision-making and, 277–284
  enforcement mechanisms, 1002
  European Convention on Human Rights, 260, 447, 863
  surveillance AI and, 797, 799
  in transnational AI governance, 256, 257
  Universal Declaration of Human Rights, 75, 863
  violations of, 176, 526, 797, 799, 891n12, 991
Huntington, Samuel, 676
Hurlbut, J. Benjamin, 984
Hussain, Waheed, 8
hybrid governance models, 575–576

I

Ibbetson, D., 299
IBM, 100n2, 155, 502, 507–509, 522, 822
ICO (Information Commissioner's Office), 284, 502, 503
ICRC (International Committee of the Red Cross), 1030, 1040n74
ICTs. *See* information and communication technologies
ideal type bureaucracy, 383, 386, 388–390, 393, 394
identity theft, 153, 864
IEEE (Institute of Electrical and Electronics Engineers), 277, 279–280, 284, 336
IHL. *See* international humanitarian law
ImageNet dataset, 346–349, 347*f*, 737
IMDRF (International Medical Devices Regulators Forum), 848
Immigration and Customs Enforcement, U.S. (ICE), 798, 800, 802, 808
immunizations. *See* vaccines
impact assessments
  algorithmic, 173, 288, 500–501, 525, 528–529
  on cybersecurity, 56
  data protection, 865
  environmental, 77–78, 495
  by European Commission, 945, 946
  in permit programs, 303, 309
  privacy-related, 467
  for public AI systems, 425, 432–433
incentive-based rating systems, 518–519, 524–530
incident response plans, 315n23
inclusivity. *See also* diversity, equity, and inclusion
  co-governance and, 997
  of healthcare, 838, 840, 845, 847, 852
  incentive-based rating systems for, 527
  institutional reform and, 993, 1003
  paradox of participation and, 985–986, 1003
  public AI systems and, 432
  question exploration tool on, 116
  representational, 983–984, 997
income
  democracy and, 715–716
  distribution of, 27, 79, 99, 710, 756–757, 759–761, 764
  economic redundancy of labor and, 768–771
  education level and, 729
  global growth of, 92, 93*f*
  for individual agents, 757–759, 757*f*
  inequality in, 27, 50–51, 722n15, 921–922, 964
  life satisfaction and, 721n3
  multiplier effect of non-labor income, 767
  near-complete automation and, 99, 711
  trust in AI and level of, 557
  universal basic, 243, 563, 656n6, 719, 742, 748, 770–771, 774nn11–12

Indigenous Data Sovereignty movement, 989
Industrial Revolution
in Big History framework, 101n4
capital–labor complementarities in, 709
colonialism and, 1003
economic impact, 38, 39, 90, 101n5
energy capture and, 91, 92, 101n6
Fourth, 988, 1002–1003
human trajectory shifts and, 91, 92, 94, 94*b*
inequality trends and, 103n22
rate of technological progress during, 87
transformative effects, 21, 23, 25, 39, 358
inequality. *See also* equality
automation and, 50–51, 54, 99, 468–469, 655, 728–729, 964, 970
digital divide, 364, 424, 538, 990
early identification of, 47
economic, 103n16, 729, 823
in educational opportunities, 210
in Global South, 717, 982, 985–986, 994, 997, 1002
income, 27, 50–51, 722n15, 921–922, 964
Industrial Revolution and, 103n22
political, 239, 987
public AI systems and, 432, 433
public opinion on, 722n15
smart city technologies and, 820–821, 823, 826
social, 225, 407, 623–624, 987
structural, 846, 1000, 1003
Infantino, M., 282
information and communication technologies (ICTs)
administrative behavior and, 387–388
coordination through, 27–28
in COVID-19 pandemic, 540
economic impact, 26–27
in Global South, 987
identity effects of, 27–28
innovation in, 27, 385
military use of, 89
power and, 28–29
in public organizations, 385, 387–391
telegraph, 26–28, 89, 888–890
Information Commissioner's Office (ICO), 284, 502, 503
informed consent, 159–160, 560, 815
Ingrams, Alex, 11, 479
injustice. *See* structural injustice
innovation
algorithmic, 102n14, 842
in commercial sector, 922, 929, 964, 972, 973
complementary, 24, 25, 86, 87, 921
Digital Innovation Hubs, 944, 948, 950, 951
drivers of, 51, 990
economic, 916
general purpose technologies and, 24–26, 86, 87, 921
globalization of, 882–886, 884*t*, 890, 891n11
in healthcare, 838, 842, 844, 852
in information technologies, 27, 385
in intelligence technologies, 29
military, 1016, 1017, 1019–1021, 1032
in nuclear weapons, 38
promotion of, 55, 158, 965
public values and, 1018
smart city technologies and, 820
Insight (adjudication software), 788–792, 794
Institute of Electrical and Electronics Engineers (IEEE), 277, 279–280, 284, 336
Institutional Analysis and Development, 394
intellectual property
copyright law and, 32, 500, 619, 773
in Global South, 990
patents, 308, 710, 780, 885, 966, 968, 990
protections for, 27, 773n6, 922, 945
trade secrets, 112, 173, 305, 467
TRIPS agreement on, 990
intelligible principle test, 465
International Committee of the Red Cross (ICRC), 1030, 1040n74
international humanitarian law (IHL), 326, 342n18, 897, 919, 927, 974, 1027, 1029
International Medical Devices Regulators Forum (IMDRF), 848
International Organization for Standardization (ISO), 336, 501
international politics, 959–975
applications of AI in, 960–963
balance of power in, 914, 924–925, 959, 971–973
civil society advocacy in, 897
implications of AI for, 923–928, 968–975

motivations for pursuing AI in, 963–967
security institutions and, 925–926
security norms and, 926–928
International Telegraph Union (ITU), 888–889
Internet of Things (IoT), 620, 800, 801
ISO (International Organization for Standardization), 336, 501

J
JADC 2 (Joint All-Domain Command and Control), 384
Janeja, Vandana, 541
Janssen, Marijn, 392, 422
Jasanoff, Sheila, 739–740, 984
Jean, Botham, 216
Jelinek, T., 402–403
Jobin, Anna, 984
Johnson, C., 557
Johnson, J. S., 974
Joint All-Domain Command and Control (JADC 2), 384
judicial review, 302, 308
Juelfs, Megan, 15, 746
Jumper, J., 617n6
justice
corrective, 134, 224
definitions of, 130, 134
DEI as, 211–212, 219, 221, 223–224
democracy and, 152
distributive, 134, 199, 205, 226
gender, 135, 141
moral framework of, 145n42
organizational, 453–454
personifications of, 153
policy strategies for, 139–142
political communities and, 235, 237
productive, 134
public AI systems and, 426
racial, 135, 138, 141, 220, 572, 595
responsibility and, 226–227
standards of, 134, 135, 138
theories of, 222–226, 228n18
justification
accountability and, 166, 173, 174
of automated decision-making, 462–464, 472–473
explainable AI and, 187, 190
of power, 204–207, 222
substantive, 205–207

K
Kak, Amba, 448, 997, 1001
Kania, Elsa B., 17, 895, 924
Kant, T., 561
Kaplan, Lyric, 619
Karabarbounis, L., 713, 750
Kelly, Kevin, 25
Kendall-Taylor, A., 970
Kenis, P., 575
Kennedy, John F., 462
Keohane, Robert O., 167
Keriwal, M., 541
Keynes, John Maynard, 648, 655
Kiaros, external audit of, 509
Kim, Kyoung-Cheol, 9–10, 383
Kim, S. H., 739, 740
Kleinberg, Jon, 679
Klievink, Bram, 11, 479
Klinova, Katya, 14–15, 726, 768
Kohler-Hausmann, I., 444
König, P. D., 482, 487
Korinek, Anton, 1, 5–6, 15, 65, 83n19, 729, 732, 746, 750, 767–768
Kovacs, A., 988–989
Krafft, P. M., 556
Kremer, M., 101n6
Kreps, Sarah E., 17–18, 959
Kroll, Joshua A., 173
Kuk, George, 392, 422
Kurzweil, Ray, 751
Kwet, Michael, 989–990
Kysar, Douglas A., 668

L
labor. *See* employment
Lai, Alicia, 11, 461
Langer, Paul F., 9, 10, 398
language models
algorithms and, 203
automation and, 51
flexibility of, 450
progress monitoring, 57
training of, 53, 57

LaPointe, C., 35
Latour, B., 486
lattice networks, 113, 113*f*
law enforcement. *See* policing
LAWs. *See* lethal autonomous weapons
Lawson, George, 889, 890
Lazar, Seth, 7, 198, 1016
Le Bui, M., 145n42
Lechterman, Ted, 7, 164
Lee, Kai-Fu, 102n15
legal issues
  accountability (*see* accountability)
  automated decision-making and, 461, 465–469, 474, 1029
  citizen suits, 307
  disparate impact bias, 131–133, 144n27, 302, 523
  disparate treatment bias, 131–133, 135, 218
  due process (*see* due process)
  fiduciary duties, 325–326, 335, 337–340, 342nn13–15
  Freedom of Information Act, 466, 479
  humanitarian law, 326, 342n18, 897, 919, 927, 974, 1027, 1029
  informed consent, 159–160, 560, 815
  intellectual property (*see* intellectual property)
  interoperability, 1025–1026, 1029–1030, 1040n75
  judicial review, 302, 308
  legal norms, 1025–1026, 1026*t*, 1028–1030
  nondelegation doctrine, 465–466
  notice requirements, 119–123, 120–121*t*
  permit programs and, 301–308
  privacy rights (*see* privacy)
  reason-giving doctrine, 467, 488
  social alignment implementation and, 79
  transparency and, 488–490
legitimacy
  of decisions, 57
  democratic, 175, 487
  of elections, 153
  of harms of AI, 995–996
  of NATO, 1027
  political, 389
  of power, 223
  procedural, 206, 207, 235
Lehr, D., 488
Leontief, Wassily, 641, 648, 654, 750
lesbians. *See* LGBTQ communities
Lessig, Lawrence, 726–728, 734, 742
lethal autonomous weapons (LAWs)
  accountability and, 897–898, 960
  bans on, 55, 79, 938, 1026, 1029, 1039–1040n73
  definitions of, 48–49, 895
  governance debates, 895–900, 910, 926
  great power competition and, 34–35
  humanitarian law and, 342n18, 897, 927, 1029
  public opinion on, 561–562, 927
  UN GGE on, 896–899, 911, 926
  use of, 49, 58n3, 561, 961
Leveson, Nancy, 442–454
Levin, P. L., 35
LGBTQ communities, 524, 595, 807
LibriSpeech dataset, 737–738
Lighthill, James, 354
Lim, Désirée, 7, 210
Lindemann, S., 575–576
Lindop Committee on Data Protection, 864–865
Lindroth, Marjo, 985
Lipsey, Richard, 87, 88*t*
List, C., 29
loss aversion, 463
loyal AI systems, 320–341
  advantages of, 321, 328–331
  alignment and, 331, 332, 341n7, 341n10
  certifications for, 336–338, 340, 341
  definitions and properties of, 322–323
  degrees of loyalty, 325–328, 341n6
  designing, 320–322, 331–333, 339
  dimensions of loyalty, 331–332
  examples of, 325–329
  fiduciary duties and, 325–326, 335, 337–340
  governance of, 334–340
  human loyalty compared to, 324–325
  policy recommendations, 339–340
  privacy and, 332, 335–336
  process-based regulation, 338–340
  product-based regulation, 337–338
  standards and metrics for, 336–341
  transparency of, 320, 328, 331–333, 335–336, 340
  value alignment and, 323, 333–334
Lumineau, F., 410
lump-of-labor fallacy, 754
Luna-Reyes, L. F., 485–488
Lynskey, O., 866

M
Maas, Matthijs M., 9, 358, 926
MacCarthy, M., 664
machine learning (ML). *See also* algorithms
  big data and, 186
  BLADE program and, 961
  bureaucratic uses of, 388
  costs associated with, 27
  in COVID-19 pandemic, 541
  cybersecurity and, 971
  cyberwarfare and, 35–36
  data analysis based on, 237
  as decision-making tool, 131, 384
  deepfake technology and, 962
  definitions of, 58n1, 198, 269n3
  disaster response and, 47
  disease outbreak warning systems and, 47
  distributive outcomes of, 240
  energy cost of, 991
  fintech and, 860, 868–871
  in healthcare, 839, 842, 846
  interpretable, 467, 869–871
  library architecture, 83n9
  natural language processing and, 402
  public policy cycle and, 536, 537
  surveillance AI and, 807
  transparency and, 429, 467
machine translation, 48, 881–883, 889–890, 899
Mackereth, Kerry, 883
Macron, Emmanuel, 1016
Mahnken, Thomas, 1019
Mahoney, Casey, 17, 914
Makhdoumi, Ali, 664, 666, 668, 684, 687–688, 691–692
Manheim, Karl, 619
Mani, G., 738, 739, 742
Marchetti, Raffaele, 993
market governance, 574–575, 764–765
Markey, Edward, 572
Marmor, Andrei, 149, 161n4
Maslow, A., 761
Matheny, Michael E., 15–16, 838
Mathews v. Eldridge (1976), 793
Matsui, Doris, 572
Mattis, James, 99
Mayer, F., 574
McAfee, A., 98
McBride, Keegan, 8, 253
McCloskey, Deidre, 101n4
McGlinchey, Stephen, 998, 999
McNealy, Jasmine, 12–13, 572
McTernan, Emily, 238
Meade, James E., 713
medical care. *See* healthcare
medical devices, 282, 300, 302, 311–313, 314n17, 847–850
Megvii, 507–508
Mejias, Ulises, 999
mental models, 448–450, 449*f*
Meta. *See* Facebook
Meyer, B. D., 760
Michaels, Guy, 921
Michel, Arthur Holland, 621
microaggressions, 227n2, 237–238
Microsoft
  collective actions against, 589*t*, 591*t*, 592, 595–596
  DialoGPT, 609
  ethical commitments from, 412
  facial recognition technology, 507–509
  in Global South, 991
  internal audits by, 495, 502
  machine translations, 881, 883, 889
  number of employees, 652, 713
  residual networks, 346–348
  workforce diversity, 527
Microsoft Research Asia (MSRA), 881–889
Milan, Stefania, 997, 999
militaries. *See also* warfare; weapons
  agent alignment and, 30
  AI governance debates, 895–911, 900–901*f*, 903*f*, 905*f*, 907–908*f*, 909*t*
  challenges of AI for, 37–38
  decision-making in, 897, 919, 965–966
  existing and emerging AI applications for, 95, 97, 98*t*
  expenditures on, 28, 34
  general purpose technologies and, 25–26
  innovation studies, 1016, 1017, 1019–1021, 1032
  loyal AI systems for, 326
  predictive capabilities, 961, 965
  social media use by, 28
  "spin on" process, 1036n21
military power, 97, 910–911, 916–924, 928, 1019–1020, 1027

Mill, John Stuart, 480
Mills, Charles W., 213, 224
Minsky, Marvin, 739
misinformation, 50, 52, 204, 539, 564, 677–679, 721n1
mission creep, 798, 801, 809, 815
Mittelstadt, B., 498
ML. *See* machine learning
modification and reproduction controls, 607–608, 611–614
Mokyr, Joel, 652
Moll, Joana, 158
Moore, J. F., 401
Moore, M. H., 434
Moore's Law, 3, 734, 752, 773n3
morality. *See* ethics
Moral Machine project, 554
Morgan, J., 299, 309
Morris, Ian, 91, 101n4
Moses, L., 299
MSRA (Microsoft Research Asia), 881–889
Müller, W. M., 410, 421, 488
multi-stakeholder convening, 109–124
  conveners' roles in, 110–112, 114
  conversation catalysts, 117–124, 120–121*t*
  knowledge foundation for discussions, 114
  motivation and experience factors, 112–113
  non-technical participants, 109–112, 114, 124
  participant wants and needs, 111–113
  question exploration tool in, 114–117
  relationship building through, 113, 113*f*
  technical participants, 111, 112, 124, 125n5
Murphy, Craig, 889
Murphy, James T., 1002
Musk, Elon, 202
Musumba, M., 47

N

Nabatchi, T., 423
National Artificial Intelligence Initiative Act of 2020 (NAIIA), 413
National Institute of Standards and Technology (NIST)
  on AI best practices, 517
  automated decision-making and, 285
  black box technology reviews by, 304, 314n20
  external audit function of, 502, 508–510
  on facial recognition technology, 508–509, 825
  permit programs and, 303, 304
  surveillance AI and, 814
  technical assessments conducted by, 355
national power. *See also* economic power; military power
  AI as tool of, 38–39, 883, 917–924, 928
  general purpose technologies and, 25, 914
national security
  AI integration in strategies for, 928–929
  inaction on AI development and, 1038n57
  permit programs and, 302, 310
  privacy violations and, 153, 154
  surveillance AI and, 484, 801, 806, 810
National Security Commission on Artificial Intelligence (NSCAI), 354, 883, 889, 1038n57
National Transportation Safety Board (NTSB), 445, 449
NATO (North Atlantic Treaty Organization), 1015–1033
  2030 agenda, 1015, 1016
  AI governance by, 1016–1017, 1021–1032, 1023–1024*t*, 1026*t*
  detention policies, 1029–1030, 1040n79
  on EDTs, 1015–1016, 1021–1022, 1027–1033
  ethics and values, 1025–1028, 1026*t*
  facilitative power of, 1022, 1029
  Framework for Future Alliance Operations, 1027
  Human View framework, 1037n44
  legal norms and, 1025–1026, 1026*t*, 1028–1030
  objectives of, 1015, 1030, 1034n1
  safety and security issues, 1026, 1026*t*, 1030–1032
  sociotechnical systems and, 1021, 1037n44
  standards and certification, 1024–1025, 1038n53
  strategic and policy planning, 1022, 1025
  technological determinism and, 1018, 1022
natural language processing (NLP). *See also* language models
  for chatbots, 422
  deepfake technology and, 962

energy cost of, 991
GLUE and SuperGLUE benchmarks for, 349, 737
Insight software and, 789
machine learning and, 402
in political communities, 236
surveillance AI and, 807
Navalny, Alexei, 718
near-complete automation, 98–99, 103n21, 103n23, 711–712, 714*t*
Nedzhvetskaya, Nataliya, 13, 584
Neiman, B., 713, 750
neoclassical production function, 642–645
Neolithic Revolution, 38, 87, 90–92, 94*b*, 101n4, 101n6, 772
networked governance, 575–577, 580–581
Newman, Andrew F., 697
Ng, Andrew, 86
Nguyen, A., 730
9/11 terrorist attacks (2001), 807, 830–833
Nissenbaum, Helen, 161n1, 166, 172, 199
NIST. *See* National Institute of Standards and Technology
Nixon, Richard, 770
NLP. *See* natural language processing
Noble, Safiya, 145n42, 517
Non Aligned Movement, 999
nondelegation doctrine, 465–466
norms. *See also* social norms
cultural, 844
gender-based, 214
legal, 1025–1026, 1026*t*, 1028–1030
military behavior, 34
mutual restraint, 36, 37
privacy, 51, 79, 161n1, 963
publication, 27, 55, 604–606, 615
security, 926–928
transnational, 926
North, D. C., 101n4
North Atlantic Treaty (1949), 1027, 1038n60
notice requirements, 119–123, 120–121*t*
NSCAI (National Security Commission on Artificial Intelligence), 354, 883, 889, 1038n57
NTSB (National Transportation Safety Board), 445, 449
nuclear weapons, 34–39, 49, 55, 65, 86, 100n1, 925, 974
nudging, 195, 203, 239, 668
Nyst, Carly, 718

O

Obama, Barack, 257, 648
Obermeyer, Z., 132, 845
Office of Appellate Operations (OAO), 781, 783–790, 792
Okin, Susan, 225
O'Neil, Cathy, 517
OpenAI, 52–53, 605, 610, 612–613
Organisation for Economic Co-operation and Development (OECD)
AI as defined by, 254–255
automatable jobs in member countries, 721n6
Council Recommendation on Artificial Intelligence, 423
ethics guidelines adopted by, 926, 1037n47
high-skilled migrants in member countries, 886
international AI policy observatory, 953
multi-stakeholder approach to AI governance, 939, 984
Principles on AI, 256–257, 276, 848, 1034n10
resistance to agendas of, 986
on sociotechnical systems, 441
on workplace automation, 562
Osborne, Michael A., 711
Ostrom, Elinor, 394
outsourcing, 54, 450, 466, 483–485, 606, 714
Ozdaglar, Asu, 677–678

P

Paine, Thomas, 720, 770
PAIS. *See* public AI systems
Palantir, 802, 826, 961, 965, 992
Paquet, M., 535
Partnership on AI (PAI), 109, 412, 615
Patent and Trademark Office, U.S., 314n21, 316n53, 316n55
patents, 308, 710, 780, 885, 966, 968, 990
pathetic dot theory, 727
Pawson, Ray, 621

peer-to-peer (p 2p) lending, 860, 863, 867, 872n2
Pelosi, Nancy, 962
Peña, Paz, 991
PERF (Police Executive Research Forum), 832
permit programs, 299–313
  accountability and, 300, 307
  advantages and disadvantages, 300, 309–310
  application review process, 304–305, 304*f*, 314n21
  application submission, 302–304, 303*f*
  determination of applicability, 301–302, 302*f*
  enforcement mechanisms, 299, 305–308
  fairness standards, 314n9
  general permission approach vs., 311–313, 312*t*
  issuance of permits, 305, 306*t*
  legal elements, 301–308
  political dimensions, 309–311
  public review process, 304–305
  transparency of, 302, 304, 311–313
personalized algorithms, 560–561, 668
personalized medicine. *See* precision medicine
Petit, Nicolas, 360
Pettit, Philip, 29, 151
PfD (AI Partnership for Defense), 928
Pichai, Sundar, 593
Pigou, Arthur, 772
Pindyck, Shira, 17, 914
Png, Marie-Therese, 18, 981
Polanyi, Michael, 644
Police Executive Research Forum (PERF), 832
policing. *See also* predictive policing
  big data as tool for, 830–833
  community models for, 831–832
  discretionary power in, 390
  facial recognition technology in, 51, 95, 202, 469, 507–509, 523, 559, 799
  gender bias in, 408
  near-complete automation of, 99
  racial bias in, 135, 145n44, 408, 523, 821, 964, 969
  surveillance AI in, 799, 827–828
  violence in, 210, 212, 216, 223
political communities, 232–246
  agency, 233, 234, 246
  AI's impact on, 235–245
  decommunitarization of, 241, 245
  democratic practices in, 235–239
  ideal of, 233–235, 238
  importance of, 232
  institutional outputs in, 235, 237, 239–240
  social ethos in, 235, 237–238, 240–241, 246
  structural changes in, 241–245
political economy, 829–833, 999–1000
political inequality, 239, 987
political power, 99, 158, 236, 681, 726, 740, 989
politics. *See also* international politics
  advertisements in, 95–96, 239
  of cognitive tools, 30
  deepfake technology and, 962
  economic impact of AI and, 714–717
  ethics and, 222
  existing and emerging AI applications in, 95–97, 98*t*
  harms of AI and, 676–681
  information technologies and, 29
  near-complete automation and, 99
  order and control in, 22, 26, 32–34, 917
  of permit programs, 309–311
  polarization in, 33, 70, 95, 676–679, 707
  voter behavior, 102n11
  workplace automation and, 562–563
Popper, Karl, 764
Povolny, Steve, 622
power, 198–207. *See also* great powers
  abuse of, 150–155, 160, 175
  accountability and, 499
  agential, 200–202
  AI in exercise of, 202–207
  algorithmic bias and, 130
  balance of (*see* balance of power)
  coercive, 167, 914, 917
  coloniality of, 1000–1002
  computing (*see* computing power)
  concentration of, 50–51, 54, 204, 207, 614–615, 972, 988
  decision-making, 236, 982, 985, 993, 995, 997
  delegation of, 165, 166, 168, 465
  discretionary, 390

economic (*see* economic power)
explanatory, 215–216
facilitative, 1022, 1029
global, 914, 960, 966, 971–972
hard, 28, 37, 258, 903, 967
information technologies and, 28–29
institutional, 1016, 1022
justification of, 204–207, 222
legitimacy of, 223
military, 97, 910–911, 916–924, 928, 1019–1020, 1027
national (*see* national power)
political, 99, 158, 236, 681, 726, 740, 989
"power to" vs. "power over," 199–200
redistribution of, 993
regulatory, 259, 990, 994
social, 233, 243, 246, 681, 701–702
soft, 258, 967
steam, 25, 87, 89–90, 94, 98–99, 101n5, 921
Power, M., 501
precision medicine, 46, 330, 891n6
predictive algorithms. *See also* predictive policing
accuracy of, 133, 135, 137, 464
background structures in, 141
bias in, 33, 129, 969
big data and, 144n24
in COVID-19 pandemic, 47, 542
for disease risk, 46, 839
in disinformation campaigns, 962
for electricity demand, 47
fairness of, 136, 442
military use of, 961, 965
in public organizations, 390
social identity variables in, 144n19
predictive policing
accuracy of, 135
data quality for, 132
fairness in, 442
gender bias in, 408
hazard analysis of, 444
heat lists and, 823, 824, 829
incentives for, 831
institutionalization of, 830
objectives of, 95
racial bias in, 408, 821, 964, 969, 992
resource allocation in, 202, 823, 833
preference orderings, 67, 72–76, 82n5, 83n13
prejudice. *See* bias
principal–agent relationships, 30, 67–69, 68*f*, 82–83n9, 166, 731
Pritchett, L., 733
privacy, 149–161. *See also* consumer financial privacy
access theories of, 149–151, 155, 161n2
automated decision-making and, 467–468
bank–customer relationships and, 861–863, 870
collective nature of, 154–155, 160
contextual, 161n1, 172
control theories of, 149–151, 155, 161n4
data collection process and, 528
erosion of, 38, 51, 149
EU's trustworthy AI and, 942
as fundamental right, 260–261
GDPR on, 254, 260–261, 332, 411, 865–868, 923
in Global South, 995–996
in healthcare, 840, 843, 850–851
HIPAA and, 630, 851
hybrid account of, 150–151, 155, 161
importance of, 152–155, 161
information technologies and, 28
loyal AI systems and, 332, 335–336
norms for, 51, 79, 161n1, 963
paradox of, 566
protection strategies, 155–161, 923
public AI systems and, 423, 432
question exploration tool on, 116
Safe Harbour Privacy Principles, 270n23
smart city technologies and, 823
surveillance AI and, 155, 797, 799, 812*f*, 814, 815, 963
tradeoffs involving, 46, 151, 485
violations of, 73, 79, 150–155, 661, 664–666, 687–689
website policies, 119, 120–121*t*, 125n13, 159
Privacy Act of 1974, 467
private nondelegation doctrine, 466
private right of action provisions, 307–308
probabilistic risk assessments, 445
Probst, J. C., 214
procedural use controls, 607, 611
process safety models, 446–447, 446*f*

productivity, 25, 87, 88*f*, 91, 920–921
productivity effect, 651, 656n3, 656n5, 672, 695
productivity paradox, 101n3
Project Maven, 593, 918, 961, 965
propaganda
  in authoritarian regimes, 970
  disinformation and, 968
  language modeling and, 51
  personalized, 154
ProPublica, 33, 134, 136, 468, 502, 509
Prosser, Tony, 364
protein folding, 46, 47, 58n2
Provan, K. G., 575
p 2p (peer-to-peer) lending, 860, 863, 867, 872n2
public AI systems (PAIS), 421–436
  accountability in, 422, 423, 428, 430–434
  codification of public values, 428–430, 434
  design elements, 423–425
  goal setting for, 425–428, 434
  governance principles, 423–435, 427*t*
  human-centeredness of, 424, 429–431, 433, 435
  impact on public values, 432–433
  as integrated governance solution, 433–435, 434*f*
  process governance framework for, 425–435, 426*f*
  public service decision-making and, 430–432, 434
  stakeholder participation, 421–425, 427–435
  transparency of, 422, 423, 426, 428–435
  value alignment and, 428
public benefits eligibility, 522–523, 526, 528
public opinion, 553–566
  on autonomous vehicles, 554
  external audits and, 503
  on facial recognition technology, 554, 558–559
  future research directions, 563–566
  importance of, 553–554
  on inequality, 722n15
  knowledge and trust as factors for, 555–558, 563–565
  on lethal autonomous weapons, 561–562, 927
  manipulation of, 407, 962
  on personalized algorithms, 560–561
  social media's influence on, 718, 721n1
  on workplace automation, 562–563
public organizations, 383–395
  administrative behavior, 383, 386–388, 390–393
  agents and agency in, 383–385, 390–392
  AI use, 383–385, 388–395, 400, 485, 491, 534–535
  authority in, 384, 392, 394, 535
  bureaucratic structure, 383, 386, 388–390, 394
  capacity building, 487–488
  communication in, 383, 386, 390, 392–395
  data set creation by, 141
  decision-making in, 384–386, 390–395
  dehumanization and loss of control in, 389, 393–395
  evolution of, 389–390, 394–395
  ICTs in, 385, 387–391
  specialization within, 388, 389, 393
  value alignment and, 383, 391–392, 394
public policy cycle, 534–545
  actors involved in, 535
  AI applications in, 537–545
  in COVID-19 pandemic, 535, 536, 539–545
  dynamic nature of, 534
  stages of, 536–537
public–private partnerships, 543–544, 801–802, 924, 948, 950, 988
public values
  codification of, 428–430, 434
  goal setting for, 425–428, 434
  impact of PAIS on, 432–433
  innovation and, 1018
  surveillance AI and, 825
  transparency and, 485–486
Putin, Vladimir, 718, 959
Pylant, A. C., 576
Pymetrics, internal audit of, 506

Q

qualified majority voting (QMV), 257, 269n13
quality of life, 46–47, 53, 54, 268, 803, 828
quasi-automation. *See* near-complete automation
queer persons. *See* LGBTQ communities

Quick Disability Determination (QDD) process, 788

R
Rachels, James, 152
racial bias
in advertising, 132, 520
of Allegheny Family Screening Tool, 844
child welfare determinations and, 522, 844
in COMPAS algorithm, 134, 136, 137, 280, 468
decision-making impacted by, 463
Emotion AI and, 521
in facial recognition technology, 50, 173, 522–523, 559, 825, 969
in healthcare, 132, 212–214, 216, 217, 839, 845–846
hiring algorithms and, 69–70, 468, 920
intrinsic racism, 228n7
in policing, 135, 145n44, 408, 523, 821, 964, 969
search engines and, 517, 518
smart city technologies and, 823–825
structural injustice and, 212–214, 216–220
in web-searching algorithms, 392
racial justice, 135, 138, 141, 220, 572, 595
Raghavan, M., 506
Raji, Inioluwa Deborah, 11, 442, 445, 495
Rakova, B., 741
Ranganathan, N., 988–989
Rasmussen, Jens, 451, 452
Raval, Noopur, 1001
Rawls, John, 223, 225, 226, 228n18, 234
Ray, Gerald K., 15, 779
Raz, Joseph, 160
Reaching Critical Will (RCW), 899
reasoned transparency, 467, 481, 484
reason-giving doctrine, 467, 488
Reddy, Raj, 735, 738
Regional Economy and Society Analyzing System (RESAS), 389
Reiner, Peter B., 8–9, 320
reinforcement learning, 50, 71, 203, 269n3, 366, 474, 964
Reitze, A., 310
responsible research and innovation (RRI), 1017–1018, 1021–1022, 1025
Restrepo, Pascual, 102n16, 645, 647, 649, 650, 653–654, 656n3, 669–673, 694–695, 729–730, 750, 921
reverse engineering, 385, 497, 622
revolutionary technologies, 25, 87, 90–94, 94*b*, 98, 888–889
Rhue, Lauren, 521
Ricaurte, Paola, 991
Richardson, R., 132
Robinson, James A., 92, 103n23, 680, 681, 701
robotic systems. *See also* automation
agency of, 67
in COVID-19 pandemic, 542, 543
employment rates and, 653
EU governance of, 938–941, 943, 952–953
as explainable AI, 188
flexibility of, 50–51
productivity and, 921
Roland, Alex, 1017
root cause seduction, 452
Rosa, Fernanda R., 989
Rosen, S., 761
Ross, C., 311
RRI (responsible research and innovation), 1017–1018, 1021–1022, 1025
Rubin, Andy, 599n8
Ruggie, John Gerard, 888
Ruhl, J., 300, 311, 313n7
Russell, Stuart, 72, 341n9, 654, 728

S
Sabin, J. E., 54
Safe Harbour Privacy Principles, 270n23
safety, 441–455. *See also* security; structured access
accident models, 444–445
accountability and, 453
of autonomous vehicles, 31, 240, 444–445, 449
of cloud-based systems, 451
control problem and, 22
control structures, 443, 443*f*, 451–452, 455
culture of, 453–454
curse of flexibility and, 450
of EU's trustworthy AI, 942
feedback mechanisms for, 452

safety (*cont.*)
hazard analysis, 441, 443–444, 443*f*, 450–451
in healthcare, 841, 842, 847, 848, 852–853
human oversight and, 448
investment in, 50, 56
Leveson's lessons for, 442–454, 455*t*
mental models and, 448–450, 449*f*
NATO issues of, 1026, 1026*t*, 1030–1032
probabilistic risk assessments, 445
process models and, 446–447, 446*f*
public AI systems and, 423, 426, 432
reporting systems, 452
social alignment on, 79
System-Theoretic Process Analysis for, 450–451
value erosion and, 23, 38
SAI. *See* surveillance AI
SaMD (software-as-a-medical device), 282, 311–313, 848
Sampath, P., 989–991, 1001
Samuel, Arthur, 519
Sanders, J. W., 166
Sandoval-Almazán, Rodrigo, 12, 534
Sandvig, Christian, 496
SBTC (skill-biased technical change), 642–645, 729
Schertzer, R., 535
schooling. *See* education
Schwartz, Paul, 260
science, technology, and society (STS) studies, 1016–1021, 1032
Scotford, Eloise, 361–362
search engines
advertising and, 204, 520
algorithms for, 172, 236, 843
bias and, 517, 520
commercial, 336
for faces, 51
loyal AI systems and, 326
secrecy, 483–485. *See also* trade secrets
security. *See also* cybersecurity; national security; policing; safety
collective, 1015
epistemic, 28
facial recognition technology and, 559
great power competition for, 34–39
international, 558, 915, 925–928, 1016, 1019, 1025, 1032–1033
investment in, 50, 158
NATO issues of, 1026, 1026*t*, 1030–1032
norms related to, 926–928
smart city technologies and, 822–823, 825–827
surveillance AI and, 813
tradeoffs involving, 37, 825
of voting systems, 174
Selbst, A., 144–145n40
self-determination, 134, 175, 205–206, 260, 865–867, 984, 989, 999
self-driving cars. *See* autonomous vehicles
self-interest, 51, 321–325, 331, 341n3, 341n8, 466, 565
sexism. *See* gender bias
sexual orientation. *See* LGBTQ communities
Shachar, Carmel, 15–16, 838
Shadbolt, Nigel, 158
Shapiro, Carl, 697
Shevlane, Toby, 13, 604
Shimshoni, Jonathan, 1019
Shoam, Y., 143n11
Siddarth, Divya, 617n7
Sidney, M. S., 536
Simon, Herbert, 383, 386–388, 390–395
Sinevaara-Niskanen, Heidi, 985
Singh, Ranjit, 982
Singhal, Amit, 739
skill-biased technical change (SBTC), 642–645, 729
smart city technologies, 820–834. *See also* facial recognition technology; predictive policing
adoption patterns, 822
Chinese, 964, 967
corporate narrative of, 822–829
critiques of, 823–829
funding for, 830, 831
political economy approach to, 829–833
real-time systems, 808
use of term, 820–822
Snowden, Edward, 34
social alignment, 69–82
AI governance and, 66, 71
challenges of, 65, 82
definitions of, 65, 69

externalities and, 65, 69–81
ideal, 65, 72–75
implementing, 77–79
importance of, 66
norms and, 72, 75–81, 80*f*
partial preferences and, 74–76
spheres of, 76–77
social ethos, 235, 237–238, 240–241, 246
social inequality, 225, 407, 623–624, 987
social insurance, 766–768, 774n9
social loafing, 463
social media. *See also specific platforms*
disaster response and, 47
disclosure rules for, 342n19
echo chambers in, 662, 676–679
military use of, 28
moderation of, 33
public opinion influenced by, 718, 721n1
recommender systems, 52, 95, 172, 203, 204
user engagement, 30, 70, 204, 239
social norms
alignment and, 72, 75–81, 80*f*
competition and, 76
fairness and, 80
informal, 78–79, 214, 234–238
in political economy, 829
privacy and, 161n1
structural injustice and, 214, 216
social power, 233, 243, 246, 681, 701–702
Social Security Administration (SSA) Disability Program, 779–794
AI use cases, 787–790
barriers to technology governance in, 785–787
eligibility requirements, 780–781
foundational infrastructure for, 782–785
historic challenges for, 781–782
Insight adjudication software for, 788–792, 794
leadership support and blended expertise in, 790–791
lessons learned from, 790–794
operational data utilized by, 791–792
Quick Disability Determination process, 788
structured policies and procedures, 783–785, 784*t*
workflow digitization and systemization, 782–783
social welfare functions, 65, 72, 74, 83nn13–14, 83n16
sociotechnical change, 358–374
AI and, 358, 361–362
barriers to regulation, 367, 370
emergence of change-centric approaches, 359–361
implementation of regulation, 370–373
material features and regulatory parameters, 366
problem logics and, 366–367, 368–369*t*, 374
rationales for regulation of, 364–365
varieties of, 362–364
sociotechnical imaginaries, 739–740
sociotechnical systems. *See also* sociotechnical change
accident models for, 444–445
constraints on, 444–445
defined, 441
governance of, 422, 1021
loyalty of, 333, 334
NATO Human View framework, 1037n44
organized complexity of, 443
in structure of society, 226
STS studies of, 1018
software-as-a-medical device (SaMD), 282, 311–313, 848
software-as-a-tool, 608–609, 616n2
Solove, Daniel, 826, 827
Soper, Spencer, 622
Sørensen, E., 575
sousveillance, defined, 175
speech recognition, 95, 237, 737–738, 773, 773n2, 886, 918
SSA Disability Program. *See* Social Security Administration Disability Program
Stanford Institute for Human-centered Artificial Intelligence, 424
Stanley-Lockman, Zoe, 18, 1015
Stapleton, Claire, 599n10
Star Trek franchise, 739–740
Stasavage, D., 101n8
steam power, 25, 87, 89–90, 94, 98–99, 101n5, 921
stereotypes. *See* bias

Stiglitz, Joseph E., 697, 750, 767
Stivers, C., 573
Stix, Charlotte, 17, 937
Stop Killer Robots campaign, 7, 561, 897, 904, 910
STPA (System-Theoretic Process Analysis), 450–451
Strandburg, K. J., 488
strategic alliances, 884–885, 884*t*
structural injustice, 210–220
  AI and, 210, 212–219
  duty of care and, 173
  economic power and, 211, 217, 220
  gender bias and, 212, 220, 224
  in healthcare, 212–214, 217
  modeling, 140–141, 143
  racial bias and, 212–214, 216–220
  responsibility and, 226, 227, 228n20
  theory of, 210–211, 214–217
structural technological unemployment, 649–650, 653
structured access, 604–616
  centralization of power and, 614–615
  cloud-based deployment, 605, 608, 610, 612–616, 616n1
  deep learning models and, 610, 611, 611*t*
  facial recognition technology and, 605
  implementation of, 609–614
  local deployment, 609–612
  modification and reproduction controls, 607–608, 611–614
  publication norms and, 27, 55, 604–606, 615
  selective information disclosure, 608–610
  use controls, 607, 610–613
STS (science, technology, and society) studies, 1016–1021, 1032
substituting force, 646–648, 650, 651, 653
Sullivan, J. X., 760
Summers, Lawrence, 651
sunk-cost fallacy, 462
Sunstein, Cass, 676, 678
supply chains
  accountability of, 266
  autonomous vehicles and, 49
  decoupling of, 35
  in Global South, 991, 996
  management of, 102n10
  for militaries, 918
  optimization of, 95
  semiconductor, 888
  transparency of, 266
Surden, Harry, 8–9, 320
surveillance AI (SAI), 797–815
  in authoritarian regimes, 22, 34, 803, 826, 962, 969, 972
  axes of, 805–810, 805*f*
  big data and, 800–801
  biometrics and, 804
  characteristics of, 799–803
  in COVID-19 pandemic, 542
  employment and, 202, 220, 622
  ethical issues, 797–798, 814, 815
  externalities and, 73
  facial recognition technology, 51, 58n4, 509, 797, 803, 962, 969
  function creep and, 799
  layering effect of, 805
  for mass surveillance, 74, 79, 542, 623, 827
  mission creep and, 798
  mitigation of harms from, 813–814
  passive vs. real-time systems, 808–809
  in policing, 799, 827–828
  privacy issues, 155, 797, 799, 812*f*, 814, 815, 963
  public–private partnerships and, 801–802
  regulatory proposals, 55, 302
  smart city technologies and, 821, 823, 825–828
  trends in, 810, 811–813*f*
surveillance capitalism, 242, 990
Susskind, Daniel, 14, 641, 655
Susskind, Jamie, 244
Sutton, R., 58n6
Sweeney, Latayna, 132, 520
system neglect, 463
System Risk Indication (SyRI), 447
System-Theoretic Process Analysis (STPA), 450–451

T

TAI (transformative AI), 23, 103n20, 358
Tan, J. S., 13, 584
Tapia, Danae, 991
task-based, capability-agnostic models, 648, 650, 653
task encroachment, 646, 647, 650–654
taxation, 246, 268, 464, 682, 732, 771–772, 990
technical bottlenecks, 96, 98, 102n14
technocracy, 174, 388–389, 682

technoglobalism, 881–882, 886, 888, 890
technological determinism, 1018–1022
technological management, 202
technological solutionism, 828, 829, 985
technological unemployment, 102n16, 407, 641, 648–650, 653, 655–656
technologies. *See also specific technologies*
AI (*see* artificial intelligence)
black box (*see* black box technologies)
dual-use, 26, 37–38, 917, 922, 1015, 1036n21
EDTs, 1015–1016, 1021–1022, 1027–1033
fintech, 521–522, 860–861, 867–871
GPTs (*see* general purpose technologies)
ICTs (*see* information and communication technologies)
persuasive, 203, 204
revolutionary, 25, 87, 90–94, 94*b*, 98, 888–889
smart city (*see* smart city technologies)
social media (*see* social media)
strategic, 25–26, 35, 37, 881
STS studies, 1016–1021, 1032
wearable, 840
technonationalism, 881–884, 887, 888, 890, 891n2
telegraph, 26–28, 89, 888–890
telemedicine, 540–542
Tesla autonomous vehicles, 445, 622
testing, evaluation, validation and verification (TEVV) framework, 56, 1031
testing and experimentation facilities (TEFs), 948, 950, 952
TFP (total factor productivity), 25, 87, 88*f*, 91
Thadaney-Israni, Sonoo, 15–16, 838
Third World Approaches to International Law (TWAIL) movement, 1001
Thomas, R. P., 101n4
TikTok, 801, 807, 887
Timaru, Ramona María, 153
Tocqueville, Alexis de, 480
Toews, R., 313n7
Torfing, J., 575
total factor productivity (TFP), 25, 87, 88*f*, 91
totalitarianism, 22, 28, 34, 89, 535
total produce lifecycle (TPLC) approach, 850
Tournier v. National Union and Provincial Bank of England (1924), 861–863
Trabucco, Lena, 18, 1015
Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement, 990
trade secrets, 112, 173, 305, 467
Trajtenberg, Manuel, 87, 96
transformative AI (TAI), 23, 103n20, 358
transgender persons. *See* LGBTQ communities
transnational AI governance, 253–268
European Union and, 253–254, 257–268
in Global South, 987–988, 995
GPAI and, 35, 256, 939, 984, 991, 1034–1035n10
OECD Principles on AI and, 256–257, 1034n10
transparency, 479–492
accountability and, 172–173
of advertisements, 262–263
AI Act on, 265–267
algorithmic, 481–491, 525, 526, 842, 846
barriers to, 482–485, 484*t*
black box technologies and, 429, 467, 484, 486, 522, 573, 846, 969
capacity-building approach, 487–490
constructivist approach, 486–487, 489–491
of data collection, 122, 572
of decision-making, 58, 466–467, 482, 483, 486–487
defined, 480, 481
democratization approach, 487, 489, 490
Digital Services Act on, 262–264
of EU's trustworthy AI, 942
explainable AI and, 481, 488
fishbowl, 466–467, 481
integrated approach, 489–492, 490*f*, 491*t*
lack of, 77, 426, 522, 606, 619, 623
legal approach, 488–490
of loyal AI systems, 320, 328, 331–333, 335–336, 340
machine learning and, 429, 467
in multinational operations, 1028
of nudging technology, 203
of permit programs, 302, 304, 311–313
in principal–agent relationships, 30
of public AI systems, 422, 423, 426, 428–435
public values and, 485–486
question exploration tool on, 117
reasoned, 467, 481, 484
structured, 608
of supply chains, 266
surveillance AI and, 810, 814
tradeoffs involving, 37, 384, 485

Trask, A., 608
Treré, Emiliano, 999
triage, regulatory, 371
TRIPS (Trade-Related Aspects of Intellectual Property Rights) agreement, 990
Trivedi, Shivangee, 622
Trolley problem, 32
Trump, Donald, 562, 962
trustworthy AI, 412, 556–558, 563–564, 941–945, 947–952
Tsang, Christine, 15, 779
Turing, A., 753
Turner Lee, Nicol, 11, 517, 520–521, 529
Tutt, Andrew, 172
TWAIL (Third World Approaches to International Law) movement, 1001
Twitter, 30, 238, 541, 561, 676, 843
Tyson, A., 557

U

Uber, 35, 449, 507, 523–524, 591*t*, 592, 921
Uighurs, 51, 58n4, 803
Ulnicane, Inga, 984
unemployment
  automation and, 54, 102n16, 649, 921, 970
  in COVID-19 pandemic, 522, 649
  fraud detection in, 522–523
  identity-mismatch and, 649
  place-mismatch and, 648–649
  skills mismatch and, 648
  technological, 102n16, 407, 641, 648–650, 653, 655–656
United Kingdom
  Brexit referendum (2016), 865, 871, 962
  changing pie effect in, 652
  duty of bank confidentiality in, 861–863, 870
  Information Commissioner's Office, 284, 502, 503
  information privacy law in, 863–867
United Nations (UN)
  Convention on Certain Conventional Weapons, 895–909, 926, 1016
  Division of Conference Management, 899
  General Assembly Voting Data, 899
  Group of Governmental Experts, 896–899, 911, 926
  Secretary General's Office of Digital Cooperation, 984
  South–South Initiative, 999
  Sustainable Development Goals, 256
  Universal Declaration of Human Rights, 75, 863
United Nations Office of Disarmament Affairs (UNODA), 898–899, 902
universal banking, 862–863, 870
universal basic capital (UBC), 720
universal basic income (UBI), 243, 563, 656n6, 719, 742, 748, 770–771, 774nn11–12
U.S. Trade Representative (USTR), 881, 889
U.S.–China relations, 881–890
  future research needs, 890
  research collaboration, 885, 886*f*
  semiconductor supply chain and, 888
  strategic alliances, 884–885
  talent circulation and, 884
  technoglobalism and, 881–882, 886, 888, 890
  technonationalism and, 881–884, 887, 888, 890
  trade practices and, 881, 889–890
use controls, 607, 610–613
utility functions, 67, 74, 82nn5–6, 83n16, 700, 760

V

vaccines, 51–52, 100n1, 541–544, 554, 808, 990
validity, external, 144n36, 471, 839
Valle-Cruz, David, 12, 534
value alignment
  defined, 72
  as element of AI control, 83n10
  loyal AI systems and, 323, 333–334
  pluralism and, 221
  public AI systems and, 428
  public organizations and, 383, 391–392, 394
value erosion, 23, 38–39
van Hoboken, J., 451
van Weel, David, 1028
VCR (Visual Commonsense Reasoning) dataset, 738
Vecchione, B., 496
Veliz, Carissa, 6–7, 141, 149
Viehoff, Juri, 7, 232

Viljoen, Salome, 243
violence. *See also* warfare
  of care, 985
  gender-based, 992
  interpersonal, 92
  9/11 terrorist attacks (2001), 807, 830–833
  in policing, 210, 212, 216, 223
  structural, 994
Visual Commonsense Reasoning (VCR) dataset, 738
Vogl, T. M., 485
von der Leyen, Ursula, 253–254, 257, 944
Vredenburgh, Kate, 6, 129

W
wages. *See* income
Waldron, Jeremy, 165, 166
Wang, Samantha X.Y., 15–16, 838
Wang, W., 282
warfare. *See also* militaries; weapons
  coalition, 1022
  cyberwarfare, 35–36, 965, 974
  escalation of, 917
  great powers and, 22, 924–925
  information, 918, 1040n74
  near-complete automation of, 99
Water, Tony, 386
Waters, Dagmar, 386
weapons. *See also* autonomous weapons; warfare
  arms control mechanisms, 26, 37, 895–896, 926
  Convention on Certain Conventional Weapons, 895–909, 926, 1016
  nuclear, 34–39, 49, 55, 65, 86, 100n1, 925, 974
wearable technologies, 840
Weber, Max, 68, 383, 384, 386–391, 393–395
Weiss, Thomas G., 986
Weissinger, Laurin B., 13, 619
Wenzelburger, G., 482, 487
Weyerer, Jan C., 9, 10, 398
Whittaker, Meredith, 599n10
Whittlestone, Jess, 5, 45
Wieringa, M., 499
Williams, Bernard, 222
Williams, Robert Julian-Borchak, 279–280
Wirtz, Bernd W., 9, 10, 398, 400, 406, 410, 421, 488
Wisconsin v. Loomis (2017), 283
women. *See also* gender bias
  gender justice and, 135, 141
  in Neolithic Revolution, 92
  public opinion on AI, 557, 565
  violence against, 992
  in workforce, 89, 224, 527
Work, Bob, 966
World Bank, 167, 761, 899
World Trade Organizations (WTO), 889, 990
Wynants, L., 543

X
XAI. *See* explainable AI

Y
Yeung, Karen, 361–362, 634
Young, Iris Marion, 212, 214–215, 217, 224
Young, Matthew M., 1, 9–10, 383, 400, 421
Younger Committee on Privacy, 864–865
YouTube, 54, 95, 668, 930n4
Yudkowsky, Eliezer, 72

Z
Zaidi, W. H., 37–38
Zang, Jinyan, 526
Zeng, J., 970
Zhang, Baobao, 1, 12, 553, 562
Zouridis, Stavros, 390
Zuboff, Shoshana, 663
Zuckerberg, Mark, 38